

**Investigating the Construct Validity of a
Concordance-based Cloze Test:
A Mixed-methods Study**

**Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester**

by

**Kunlaphak Kongsuwannakul
School of Education
University of Leicester**

2017

Abstract

Investigating the Construct Validity of a Concordance-based Cloze Test: A Mixed-methods Study

Kunlaphak Kongsuwannakul

This study has two primary goals. The first is to develop a concordance-based cloze test (henceforth ConCloze), the process of which can be divided into seven stages, ConCloze 1–7. The second goal is to investigate the validity of score interpretations, which can be categorized into five aspects according to Messick's (1995) model of construct validity: content, substantive, structural, generalizability, and external. By validity, Messick (1989: 13) referred to an integrated judgment of adequacy and appropriateness of construct-related inferences based on, e.g., item responses, observations. The sampling methods are convenience and snowball samplings, seeking non-native English speakers of mixed backgrounds in first language, who are studying in or have graduated from higher education. The analytical measures include reliability analysis, verbalization analysis, usability testing, correlation analysis, content analysis, regression analysis, Rasch modeling, sample analysis, and thematic analysis.

The construct domain is found to be very likely composite: at least lexical-semantic knowledge, knowledge of synonymy, knowledge of collocation, knowledge of grammatical structure, world knowledge, knowledge of word association, knowledge of semantic prosody, and knowledge of individual and compositional lexical-semantic content are tested. The test purpose is initially set to be a proficiency test on professional and academic English grammatical and vocabulary use, which is eventually refined in light of empirical findings in the investigation into a proficiency test on academic English vocabulary use, with the primary domains of knowledge of lexical-semantics and knowledge of word association. The subdomains involved become knowledge of core components in word meaning, knowledge of individual and compositional lexical-semantic content, knowledge of collocation, and knowledge of semantic prosody, with world knowledge, synonymy knowledge, and knowledge of grammatical structure functioning as construct-peripheral. Judged by the adequacy and appropriateness of response and score interpretations, an integrated evaluation is that the construct inferences for the ConCloze item type have validity.

Acknowledgments

Biggest thanks ever go to my supervisor. Your continuing support and patience have made my doctoral journey not so much thrown in the deep end. While there was tundra and beast-infested jungle along the way, your advice is a torch piercing through the darkness.

My gratitude also goes to Nick, Agneta, Professor Schmitt, and Dr. Shaw. It is never an easy task to produce a long piece of writing with my non-native English, but all the feedback means pushing my boundaries of potential.

To all the University staff, whose functions have played important roles at varying stages of my studies. To all the participants, whose responses made the analyses possible in an impressive way.

Gratitude is for the Royal Thai Government scholarship. Without it, studying in the UK would have been simply out of my financial reach. Also to the ETS Doctoral Small Grant (Dec 2014), which made the completion much more pleasant. Last but not least, to my family, whose love cherishes my world.

Table of Contents

Abstract	ii
Acknowledgments.....	iii
Table of Contents	iv
List of Tables	vii
List of Figures	ix
List of Abbreviations	xii
Chapter 1 Introduction	1
1.1 Rationale	1
1.2 Background and Originality	4
1.3 Research Questions and Hypothetical Construct	16
1.3.1 Research questions	16
1.3.2 Hypothetical construct.....	19
1.4 Potential Impact	24
1.5 Ethical Considerations.....	26
1.5.1 Anonymity and confidentiality.....	26
1.5.2 Incentive use.....	26
1.5.3 Copyright in concordance texts	27
1.6 Concluding Remarks and Thesis Structure	27
Chapter 2 Literature Review	30
2.1 Introduction	30
2.2 Concordance Use in Language Assessment.....	31
2.3 Framework for Construct Measurement.....	33
2.4 Lexical Priming in Cloze Procedure	54
2.5 Concluding Remarks	61
Chapter 3 Developing ConCloze	65
3.1 Introduction	65
3.2 ConCloze 1: Quantitative Prototyping	67
3.2.1 Rationale.....	67
3.2.2 Test specification (spec).....	69
3.2.3 Population and sample.....	85
3.2.4 Test responses.....	89
3.2.5 Textual feedback	98
3.2.6 Conclusion and decision.....	102

3.3	ConCloze 2–4: Qualitative Prototyping	105
3.3.1	Rationale.....	105
3.3.2	Test spec	106
3.3.3	Testing usability	109
3.3.4	Substantive content.....	124
3.3.5	Conclusion and decision.....	145
3.4	ConCloze 5: Field-testing.....	147
3.4.1	Rationale.....	147
3.4.2	Test spec	149
3.4.3	Test responses.....	157
3.4.4	Decision study	173
3.4.5	Summary	176
3.5	Concluding Inferences and Decision.....	177
Chapter 4	Using and Applying ConCloze.....	185
4.1	Introduction	185
4.2	ConCloze 6: Operational Use.....	186
4.2.1	Rationale.....	186
4.2.2	Test spec	187
4.2.3	Test responses.....	195
4.2.4	WAF–ConCloze relationship	216
4.2.5	Inferences and decision	224
4.3	ConCloze 7: Substantive Fine-tuning.....	226
4.3.1	Rationale.....	226
4.3.2	Test-taking processes and strategies.....	228
4.3.3	Substantive contrast of sample	235
4.3.4	Modified constructed-response format.....	246
4.3.5	Concluding claims and decision.....	256
Chapter 5	Concluding Remarks	260
5.1	Summary and Discussion	260
5.2	Research Limitations.....	275
5.3	Recommendations for Future Research	283
5.3.1	On language testing	283
5.3.2	On language acquisition and pedagogical practices	286
Appendices	289
Appendix 1. ConCloze 1 Items	290
Appendix 2. ConCloze 2–4 Items	329

Appendix 3. ConCloze 5 Items	334
Appendix 4. ConCloze 6 Items	364
Appendix 5. ConCloze 7 Items (Modified Constructed-response, Section 4.3.4)	388
Appendix 6. Verbal Reports for Processing Analyses in ConCloze 2–4 and 7 (Sections 3.3.4 and 4.3.2)	390
Appendix 7. Verbal Reports for ConCloze 7 (Sections 4.3.2 and 4.3.3)	408
Appendix 8. Verbal Reports for Modified Constructed-response Processing Analyses in ConCloze 7 (Section 4.3.4)	479
References	484

List of Tables

Table 1 Operational research questions	16
Table 2 Works at UCLES related to concordancing and concordances.....	32
Table 3 Aspects of knowing a word receptively (Nation 2001: 27, adapted)	34
Table 4 Areas of potential improvement.....	62
Table 5 From ConCloze's background to development.....	65
Table 6 Guiding language for ConCloze 1	71
Table 7 ConCloze 1 target words (keys).....	75
Table 8 AVL in contrast with some other word lists	77
Table 9 Gender of ConCloze 1 respondents	87
Table 10 Age of ConCloze 1 respondents.....	87
Table 11 Education levels of ConCloze 1 respondents.....	88
Table 12 L1s of ConCloze 1 respondents	88
Table 13 Length of staying or living in English-speaking countries	89
Table 14 Year of latest standardized English test	89
Table 15 Alphas-if-items-deleted in ConCloze 1.....	91
Table 16 ConCloze 1 item difficulty and discriminability.....	94
Table 17 Responses to items of good qualities	97
Table 18 ConCloze 1 respondents' comments.....	100
Table 19 Major evidence and inferences	102
Table 20 Guiding language for ConCloze 2.....	107
Table 21 Demographic profile of ConCloze 2–4 respondents.....	112
Table 22 A revision to the guiding language of ConCloze 3ff.	119
Table 23 Item responses and respondents	129
Table 24 Major evidence and inferences	145
Table 25 Guiding language for ConCloze 5.....	150
Table 26 Expected effects of modification on ConCloze 5	152
Table 27 Pairwise correlations between ConCloze and numerical demographic variables.....	161
Table 28 Linear modeling between scores and categorical demographic variables	166
Table 29 Facility, discrimination, and alphas-if-item-deleted of ConCloze 5 items	167
Table 30 Major ConCloze 5 evidence and inferences.....	176
Table 31 Evaluating the operational research questions	180
Table 32 Major evidence in pursuit of adequacy	181
Table 33 Major cases in pursuit of appropriateness.....	183

Table 34 Argument for ConCloze validity.....	185
Table 35 Guiding language for ConCloze 6.....	188
Table 36 Target words and distractors in ConCloze 6.....	190
Table 37 ConCloze 5–6 anchor items	191
Table 38 Pairwise correlations between scores and numerical demographic variables.....	198
Table 39 Tests of between-subjects effects on ConCloze 6 scores.....	202
Table 40 ConCloze 5–6 concurrent item measures and fit statistics.....	209
Table 41 Pattern matrix of IPs and extracted PCs.....	222
Table 42 Major ConCloze 6 evidence and inferences.....	225
Table 43 Details of ConCloze 7 respondents.....	249
Table 44 Major ConCloze 7 evidence and inferences.....	257
Table 45 ConCloze-score interpretations with validity aspects emphasized	261
Table 46 Evaluating the operational research questions	263
Table 47 Cases for appropriateness of score interpretations.....	268

List of Figures

Figure 1 Components of a ConCloze item.....	1
Figure 2 Messick's (1988: 42, adapted) model of test validity	2
Figure 3 Concord Writer's (Cobb 2013) concordance lines	5
Figure 4 Stevens's (1991a: 38) concordance-based vocabulary exercise	7
Figure 5 Butler's (1991: 36) prototype item	8
Figure 6 AntConc's (Anthony 2014) concordance view	10
Figure 7 Hargreaves's (2000: 211) concordance-based cloze item	12
Figure 8 CELA's (2010b) sample item in CAE Use of English	14
Figure 9 CELA's (2016a; 2016b, adapted) 'can-do' construct of the CAE program	21
Figure 10 Hiding the KWIC in an AntConc concordance	25
Figure 11 Measuring the construct proficiency with the ConCloze item type.....	37
Figure 12 Example of person-item relationship in ConCloze testing.....	40
Figure 13 A WAF item (Read 1998, cited in Cobb ca. 2011, adjusted)	48
Figure 14 Gyllstad's (2007; 2009) COLLEX and COLLMATCH.....	53
Figure 15 Example of a cloze test.....	54
Figure 16 Cloze procedure in ConCloze processing.....	56
Figure 17 Lexical priming in ConCloze processing.....	59
Figure 18 A ConCloze sample item.....	70
Figure 19 ConCloze 1 sample item (Item 0).....	73
Figure 20 ConCloze 1 Item 1	81
Figure 21 Participant number of ConCloze 1.....	86
Figure 22 ConCloze 1 comment boxes	99
Figure 23 ConCloze 1 respondents' comments on test length.....	101
Figure 24 ConCloze 1 respondents' comments on item difficulty.....	102
Figure 25 A ConCloze 2-4 Item 2	108
Figure 26 A ConCloze sign-up page, with representative examples	111
Figure 27 ConCloze 2-4 Item 1	115
Figure 28 Comprehending the test task.....	116
Figure 29 Potential incomprehension of the nature of the concordance	117
Figure 30 ConCloze 3 sample item.....	120
Figure 31 Respondents' comments on font size.....	121
Figure 32 ConCloze 4 directions for adjusting the font size.....	122
Figure 33 Respondents' comments on item presentation.....	123

Figure 34 Session of eliciting verbal reports.....	125
Figure 35 Bouts of retrospection during concurrent verbalization.....	127
Figure 36 Verbalization of meaning-compatibility check.....	130
Figure 37 Verbalization as manifest ConCloze competence.....	132
Figure 38 Tagging processes in verbal reports.....	133
Figure 39 Analyzing verbal reports for processes.....	134
Figure 40 Process ‘Testing compatibility of a given word in context’	136
Figure 41 Strategy ‘Focusing on clue-containing parts’	138
Figure 42 Process ‘Rationalizing word combinations’	140
Figure 43 Process ‘Recognizing word associate(s)’	142
Figure 44 Strategy ‘Assessing item components and difficulty’	144
Figure 45 Test-taking processes and strategies in ConCloze 2–4 Item 2.....	145
Figure 46 Transforming a ConCloze 1 item to a ConCloze 5 one	151
Figure 47 Collecting self-ratings of English proficiency	155
Figure 48 ConCloze 5 introduction page	156
Figure 49 ConCloze 5 participant number	157
Figure 50 ConCloze 5 score distribution	159
Figure 51 Respondents’ feedback on suitability of ConCloze 5 difficulty	160
Figure 52 ConCloze 5 and regression standardized residuals.....	163
Figure 53 ConCloze 5 average IFs compared with precursor ones in ConCloze 1.....	169
Figure 54 ConCloze 5’s reliability in simulated test truncation.....	175
Figure 55 Difficulty effects on test-task content of ConCloze.....	179
Figure 56 ConCloze 6 sample item.....	194
Figure 57 Number of participants across test battery.....	196
Figure 58 Score statistics and distribution of the test battery and ConCloze.....	197
Figure 59 ConCloze construct amidst self-ratings	199
Figure 60 ConCloze 6 and standardized regression residuals.....	201
Figure 61 Estimated marginal mean score and education level in L1 groups.....	203
Figure 62 Equating ConCloze 6 to 5.....	205
Figure 63 Item estimates of ConCloze 5 and equated ConCloze 6.....	206
Figure 64 ConCloze 5 Item 13 responses	211
Figure 65 ConCloze 5 Item 3 responses	212
Figure 66 ConCloze 6 Item 23 item responses	213
Figure 67 ConCloze 6 frequency levels and ascending item estimates	215
Figure 68 Scatterplot of ConCloze 6 and WAF total scores	217
Figure 69 Analyzing principal components of ConCloze–WAF item responses	220

Figure 70 ConCloze 6 and WAF eigenvalue Monte Carlo simulation	221
Figure 71 Investigating substantive processing for a generalizability argument	227
Figure 72 Enumerating test-taking processes and strategies.....	230
Figure 73 Strategy ‘Focusing on clue-containing parts’ and process ‘Choosing an action or solution suitable to the situation in hand’	232
Figure 74 Process ‘Taking in context information’	234
Figure 75 Contrastive analysis with face-validity evidence.....	236
Figure 76 ConCloze 2–4 Item 4 with a collocation list.....	237
Figure 77 Potential clues in Item 4	239
Figure 78 Conceptualized ConCloze processing	241
Figure 79 Halle’s phrasal checks	242
Figure 80 Igor’s retrospection	244
Figure 81 Igor’s phrasal checks	245
Figure 82 ConCloze competence based on a contrastive sample.....	246
Figure 83 Modified constructed-response format: From administration to analysis	247
Figure 84 A two-part ConCloze item.....	248
Figure 85 Intervention with an option sheet, and answers attempted	250
Figure 86 ConCloze 2–4 and 7 processes and strategies in number	251
Figure 87 Comparing ConCloze 2–4 and 7 processes and strategies.....	252
Figure 88 New processes in ConCloze 7	254
Figure 89 Potential processes and domains of knowledge identified for the proficiency testing	264
Figure 90 Spec-driven substantive investigation	270
Figure 91 A construct of the ConCloze item type.....	272
Figure 92 Schmitt et al.’s (2004a: 59) receptive format for a formulaic sequence	283

List of Abbreviations

AFL	Simpson-Vlach & Ellis's (2010) Academic Formulas List
AKL	Paquot's (2010) Academic Keyword List
AVL	Gardner & Davies's (2014) Academic Vocabulary List
AWL	Coxhead's (2000) Academic Word List
BERA	British Educational Research Association
CAE	Certificate in Advanced English, Cambridge English: Advanced
CELA	Cambridge English Language Assessment
COCA	Corpus of Contemporary American English
ConCloze	Concordance-based cloze test
ETS	Educational Testing Service
GSL	West's (1953) General Service List
ID	Item discrimination index
IELTS	International English Language Testing System, a test program by CELA
IF	Item facility index
IP	Item parcel
IV	Item variant
KWIC	Key-Word-In-Context
L1	First language
SD	Standard deviation
TOEFL	Test of English as a Foreign Language, a test program by ETS
UCLES	University of Cambridge Local Examinations Syndicate
UMS	Unweighted mean-square
WAF	Read's (1993; 1998) Word Associates Format
WMS	Weighted mean-square

Chapter 1 Introduction

1.1 Rationale

As in all language testing... it is crucial to make sure that all attempts to innovate in test design... by involving corpora are fully justified. This might occur by offering the possibility of meeting a need that has not yet been met,... [or] by offering the possibility of assessing some aspect of language behavio[r] or proficiency that could not be measured before (Alderson 1996: 257–8).

As suggested above, a corpus-based test design that is innovative needs a sound basis. Given the need, this study has two major goals: to innovate a corpus-based test design and to justify that innovation. The topic of test innovation and research originality will be explored extensively in Part 1.2. In this part, the need to justify the interpretations of scores from concordance-based cloze testing will be dealt with. For convenience, I name the test *ConCloze*, illustrating its composition (CONcordance-based CLOZE test). A sample item with three basic item components—prompt, stem, and options—is depicted in Figure 1 (answer: *B finish*). Over the test-development process, the components will be improved on the basis of empirical responses and will be presented subsequently.

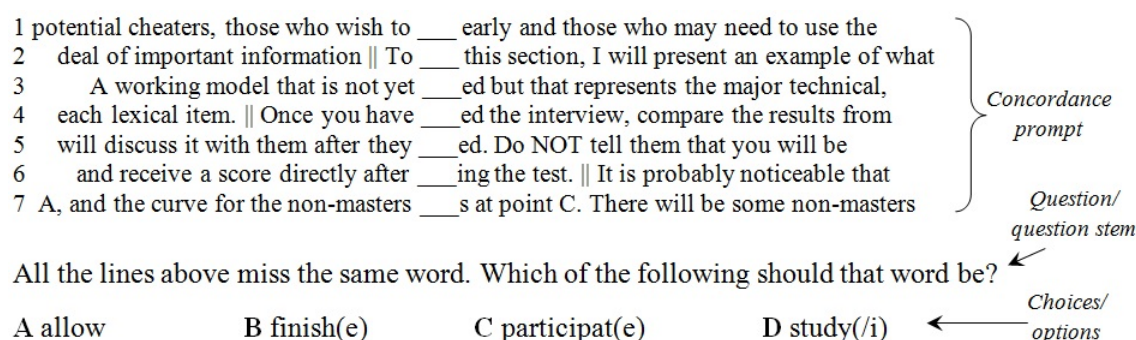


Figure 1 Components of a ConCloze item

Based on the premise of ConCloze innovativeness, this study adopts Messick's (1989) model of validity as the most appropriate to investigate the validity of this item type. A unified concept of construct validity, the model is deemed one of the most comprehensive validity theories available presently (Bachman 2000; Brown 2000; Kane

2006; McNamara 2006; Fulcher & Davidson 2007; Moss 2007; Rigney *et al.* 2008; Kane 2012a). Construct validity refers to an integrated evaluation of the extent to which particular inferences can be deemed adequate and appropriate based on test scores and other evidence (Messick 1989: 13). Accordingly, investigating validity can be equated to defining a score meaning.

Messick's model views construct validity as the indispensable element of any test interpretation and use. Illustrated in Figure 2 below, the indispensability is to the degree that giving a proper meaning to the test scores is pivotal to any validity argument. Construct validity prevails and thus encompasses all other aspects of any one test such as its utility and perception of its underlying values. For this reason, it is of paramount concern for all stakeholders involved. In establishing construct validity, validation is a major inquiry process seeking the 'holy grail' through measurement and research (Cronbach & Meehl 1955; Gabrenya 2003). Making appropriate interpretations of test scores can be deemed part of the main agenda for modern-day language testing (cf. Bachman 2000 for contemporary agenda of language testing). All these reasons indicate that investigating the validity of ConCloze-score interpretations is of utmost significance in its own right. Given this indispensability, the present study seeks to investigate construct validity for the item type by collecting empirical evidence as well as providing theoretical grounds warranting the model of its score meaning.

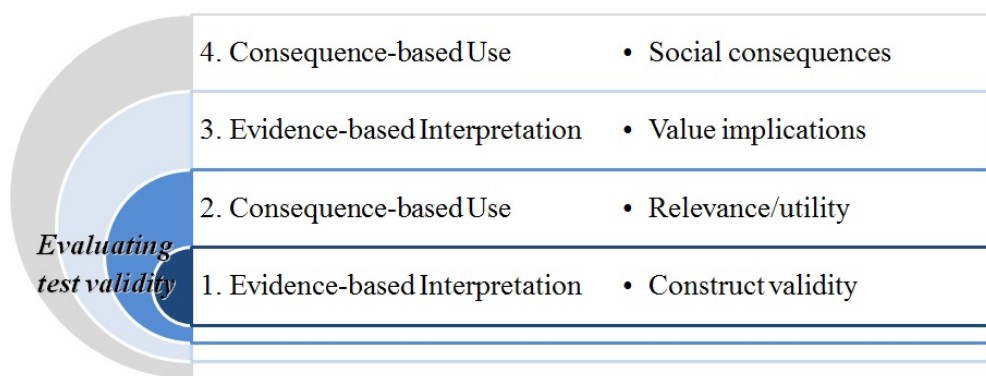


Figure 2 Messick's (1988: 42, adapted) model of test validity

In addition to the significance of investigating a construct, focusing on the score interpretations for the ConCloze item type can also be useful on an industrial scale. Testing organizations such as Cambridge English Language Assessment (CELA,

responsible for the testing program IELTS – the International English Language Testing System, for instance) and the Educational Testing Service (ETS, responsible for TOEFL – the Test of English as a Foreign Language, for instance) constantly strive to develop and improve their tests. One way of doing so is through task design. With this study validating a score meaning that aims for a power of generalization (to be discussed in Part 1.3), the testing organizations could use the meaning as a building block for interpreting the scores when developing their own ConCloze items. Having a validated meaning in hand offers them a more versatile tool for such improvement than, for example, obtaining a set of reliable ConCloze items that are ready-made but whose score meaning might not be applicable to their newly developed items. This implies that while it could be useful to come up with such a finite set of items, focusing on the central meaning that underlies the scores for those items can be even more useful considering the newness of the current item type and the potential to apply it for the benefits of the testing industry. On this account, a second reason for investigating the construct validity is the chance of wide utility for industrial test improvement (cf. also pages 16f. for the line of inquiry into the central meaning of the ConCloze scores).

Apart from the usefulness in industrial testing, emphasizing score interpretations in this study also has a psychometric merit. Multiple test methods and test-task formats almost always bring about greater fairness for the examinees (Galaczi & Khalifa 2009; Powers 2010). For example, some examiners may be more apt to perform better in one test format whereas the others could deal better with another form of test questions measuring the same competence. Providing a validated meaning for the ConCloze item type would mean that test developers can use the meaning in, for instance, selecting an alternate item type for the same domain of competence and swapping it for an in-house built ConCloze. Their assessment program would then measure the target competence more fairly to the benefit of all the test takers. Accordingly, the third reason supporting the present study is that it introduces a possibility of fairer assessment in testing practices. In sum, the present study's emphasis on ConCloze score interpretations is significant in its own right, has broad utility for the testing industry, and could help deliver fairer assessment in language testing.

1.2 Background and Originality

In the previous part 1.1, the necessity to investigate the validity of ConCloze-score interpretations has been stressed. In this part, the item type will be distinguished from previous concordance-based cloze tests, thereby establishing its innovativeness. The aim is to show that the format of a ConCloze item as it stands has never been developed and systematically administered on a large scale. Background to the item type will also be set along the discussion, which is structured as follows. First, some background regarding one of the major components of ConCloze, the concordance, will be given. Then the differences ConCloze has when compared with the previous tests will be highlighted. Finally, the originality argument will be summarized.

A corpus is a large collection of authentic texts that are digitally stored (Taylor & Barker 2008). A concordance is a user-interface platform, usually referring to “a [corpus-derived] list of contexts exemplifying a word or word family” (Nation 2001: 111). An example is given in Figure 3 below, where the search query *construct* is displayed in the node position amidst its contexts. The node position is usually called the Key-Word-In-Context (KWIC), which this study replaces with a blank in each line of the item prompt (cf. a sample item, page 1). For clarity, this study uses the word *co-text* when referring exclusively to a string of text surrounding a gap or the KWIC position in a concordance. And for convenience, the word *context* will be used for a linguistic and/or extralinguistic context in general and in relation to the KWIC position in particular (cf. Tognini-Bonelli 2001 for more specialized meaning of these two terms).

Concord Writer v.4.2
(User's text is input)

Produce real English with models from 2000
co-writers

* In French, Turkish &
汉作-版-面-世-！

for definitions o
(English-> Engl
Mandarin Chint

Concordance for *family* CONSTRUCT in **acada_corp.txt** sorted Keyword

Dictionary
Eng_Eng
Thesaurus

Click any **KEYWORD** for more context

001. ☐ ng that the potential for place lies unexamined. It is the purpose here to **CONSTRUCT** a theology of architecture that locates itself more usefully in r

002. ☐ res, and (3) opinion on the performance of the 37 IGH design features, was **CONSTRUCTED** for a pilot study with 299 respondents and subsequently the mai

003. ☐ professionals should work together to incorporate landscape views into new **CONSTRUCTION** projects and existing facilities in order to provide the patie

004. ☐ the impact of World War I on British society through politics, design, and **CONSTRUCTION** of state-funded housing in Birmingham, England during the inte

005. ☐ organizations, the US Government has authorized funding for the design and **CONSTRUCTION** of a building that serves as a center for women's leadership i

006. ☐ a close investigation of the late twelfth-century architectural design and **CONSTRUCTION** in northern Burgundy, this dissertation investigates the role

007. ☐ at Chablis, and regional monuments, this study reveals the consistency in **CONSTRUCTION** of a group of Early Gothic churches in a microregion of the Yo

008. ☐ rms, the workers traveled to northern France where they designed and began **CONSTRUCTION** on the Cistercian church of Châ alis, daughter of Pontigny. B

009. ☐ sons and sculptors combined Cistercian tendencies with techniques of local **CONSTRUCTION**, the Cistercian elements have been difficult to isolate. Appro

010. ☐ c churches with elements consistent with current trends in Å le-de-France **CONSTRUCTION**. These churches underscore the sophistication of Early Gothic

011. ☐ construction. These churches underscore the sophistication of Early Gothic **CONSTRUCTION** south of Paris. Communities of place are in need of a design p

012. ☐ basic core that is designed to support future expansion through self-help **CONSTRUCTION**. The challenge is to provide a higher density model for self-h

013. ☐ truction. The challenge is to provide a higher density model for self-help **CONSTRUCTION** that replaces the predominant scheme of the single house per l

014. ☐ of S. Lucia, which she built in her hometown of Grottammare in the Marche. **CONSTRUCTION** of the cathedral at Lincoln began during the tenure of Bishop

015. ☐ . With Canterbury Cathedral, Lincoln represents the beginning of an era of **CONSTRUCTION** which in the number and scale of its buildings has few paralle

016. ☐ to enable smooth exploration of sketched form by point-dragging. The user **CONSTRUCTS** a high-level proxy description that can be used, somewhat like a

017. ☐ beginning of the fourth century? How should its Maxentian physical form be **RECONSTRUCTED**? What evidence is there for its original function? What desig

018. ☐ p the designer move directly from existing 2-D sketches to 3-D models. Our **RECONSTRUCTION** techniques generate a novel kind of 3-D model which maintain

019. ☐ f Florence, and the Ospedale degli Innocenti in Florence has allowed for a **RECONSTRUCTION** of the process by which Suor Domenica was able to build her

020. ☐ l monograph exists. I offer a description of archaeological evidence and a **RECONSTRUCTION**, supported by measured drawings and other visual documentati

021. ☐ s, account books, diaries, letters, notebooks, fiction and non-fiction and **RECONSTRUCTS** each author's life to provide an architectural biography of th

Figure 3 Concord Writer's (Cobb 2013) concordance lines

In conjunction with other techniques, concordances have been widely used in linguistic analyses (e.g., Sinclair 1991; Flowerdew 1996; Partington 1998; Choi *et al.* 2003; Aston *et al.* 2004; Charles 2007). In pedagogical contexts, a primary use of concordances is to promote vocabulary learning (e.g., Thurstun & Candlin 1998; Aston 2002). A first example is Cobb (2013), which is illustrated in Figure 3 above. A web-based tool, Cobb's concordance-based platform is called Concord Writer, claimed to allow students to check multiple co-texts of a KWIC on demand and at their own pace, assisting in their vocabulary use for writing, for instance. Another example is Gaskell & Cobb (2004), who recommended based on their preliminary results that concordances be inserted into learners' texts as an effective learning tool for addressing writing errors. The last example is Stevens (1991a), who recommended a concordance-based classroom exercise. Illustrated in Figure 4 below, Stevens argued that, unlike traditional cloze exercises, the material does not suffer domino effects (i.e., getting one wrong answer leading potentially to getting the others wrong; cf. Figure 15, page 54 for an example of a short cloze test). He contended that this can help to boost the students' confidence because of a higher success rate in dealing with the exercise format than with vocabulary exercises in general.

Here are some words used in a biology lecture during the 5th week of a first-year biology class:

anchor	excess	obstacles
attached	extend	obtain
certain	hierarchy	oral
classification	entangle	coil

Below, you find the result of a "concordance" made on some of these words. In this concordance, a computer looked at all the readings in the first-year biology workbook. Then the computer printed each line containing those words. (The computer doesn't know where words or sentences begin or end; it just prints the line.)

DIRECTIONS: Replace each BLOCK of blank spaces below with ONE WORD from the word list above.

- 1a make up the taxonomic .
 b one progresses down the , the number of organis
 c At the bottom of the the differences within
- 2a s a longitudinal layer ing the length of each segme
 b to form one cord, which s along the length of the
 c single large taproot deep into the soil with oth
 d which the root hairs . Inside the epidermis is
- 3a Numerous granules are to the matrix side of the
 b ividual cells, firmly to each other, rest on a
 c , by which muscles are to bones, are composed
- 4a to capture prey or to the organism in place.
 b airs of chaetae. They each segment in the soil.
 c ecause roots are the ing and absorbing organs of t
- 5a tractile vacuole removes water from the cytoplasm of
 b The epidermis prevents ive water loss and yet al
 c ncreases the chances of ive water-loss but this is pr
- 6a ng to the cells. In parts of the cell the ER i
 b ng to the cells. In animals, cilia covering th
 c Organisms which have basic features in common ar
 d e, the mouse develops symptoms and dies. Howe
- 7a ilia sweep food into an groove on the side of the cel
 b side of the cell. The groove leads to the cytophar
- 8a he science of biological is known as taxono
 b is the largest unit of . It is split into
 c The various units of - kingdom, phylum,

Figure 4 Stevens's (1991a: 38) concordance-based vocabulary exercise

According to Nation (2001), an explanation for benefits of concordances is rich information contained in their authentic language, particularly multiple aspects of linguistic information about the node word such as parts of speech, word forms and affixes, collocates, referents and semantic prosodies, and grammatical patterns (e.g., Sinclair 1991; Hoey 2000; Woolard 2000; Hargreaves 2000 for examples). Discoverable primarily because of their format (to be detailed later), these aspects of information are usually of routine patterning with the KWICs and may be called multi-faceted information about them (cf. also page 64 for different terms used to refer to multicomponential information related to the KWIC in each ConCloze task). Elsewhere in this thesis, such benefits will be brought up as a unique distinction that makes

ConCloze innovative and enables the item type to deliver at least part of the benefits in the testing context. As such, it is worth emphasizing that the multi-faceted information about a KWIC refers to any information that can be found co-occurring in its contexts and that usually goes beyond the meaning of the KWIC word itself. Thus, *word information* will be used in this study in referring to the similar information with an emphasis on that related to a word, particularly the word required in the KWIC position of a ConCloze task.

In addition to briefing the background to concordance use, existing studies into concordance-based cloze testing will also be explored in this part. The aim is to locate the originality of the current research. There are three previous works in language testing identified as pertaining to this issue. They are closely related to one another and will be discussed in chronological order. Cited by Flowerdew (1996) and Hargreaves (2000), the earliest identifiable is Butler (1991). Butler reported a test-prototyping program for the then University of Cambridge Local Examinations Syndicate's (UCLES) Certificate in Advanced English (CAE). His program was performed on a small group of native speakers, and post-intermediate and advanced non-native respondents. The prototype had 13 items and was named "concordance generated cloze test" (ibid.: 29). Butler claimed the test was both innovative and authentic, fulfilling UCLES's requirements at the time for test-task design. For discussion purposes, one item is illustrated in Figure 5 (answer: *still*).

Each of the sentences below has the *same* word missing.
Fill in the blank with the correct word.

A.

1. Fortunately we have large amounts of exploitable potential on which to capitalize.
2. There is no question, however, that food production will have to be raised higher to help feed the world's growing population.
3. This does not solve the problem.
4. Here's hoping you're in your old flat by the time this letter reaches you.

Figure 5 Butler's (1991: 36) prototype item

The CAE items have three–four complete sentences each. Exemplified in Figure 5, all of the sentences require the same word to fill out their blanks. The sentences are left-aligned just like ordinary texts, not centrally-aligned as those in Figure 1 (page 1) are

in the current study. This distinction in sentence alignment could be profoundly significant to the present discussion in two interconnected facets: the information presented, and information processing.

Regarding the information presented in a concordance, Figure 5 (Butler's) is worth comparing with Figure 1 (ConCloze, page 1). Figure 5 presents full sentences as the item prompt whereas Figure 1 presents the texts in the form of truncated lines. If Butler's prompt and ConCloze's prompt had the same number of texts, then Butler's is likely to have several more words per item than ConCloze is. This could be a marked difference and potentially signify that the amounts of information imparted to the examinees, when everything else is equal, are unlikely to be equal in the two tests.

Given the possibility concerning imparting unequal information, a construct-related implication could potentially arise: differences in what could be processed during test-task engagement (cf. also Kongsuwanakul 2014a; Kongsuwanakul 2014b; Kongsuwanakul 2015b for theoretical discussion on ConCloze processing). Normally, a sentence view of concordance lines—i.e., an extended view of co-texts for each keyword search, usually in an entire sentence—is deemed less challenging than the traditional concordance view with truncation (Tribble 2013). Accordingly, it may be inferred that Butler's format could be less difficult than ConCloze. Considering the differences in information load and in item difficulty, an argument is that their constructs are unlikely to be identical.

Figure 6 below displays a concordance with truncated concordance lines. This is a standard presentation format and can be found default-generated by concordancers, e.g., WordSmith Tools (Scott 2012), SARA (BNC Consortium 2005) (cf. Sinclair 1991; Cobb 1997; Sinclair 2003; Teubert & Čermáková 2004 for more examples). As far as a proper designation is concerned, users' and learners' first encounter with a concordance is likely to be of a truncated one. By contrast, Butler's format does not feature a truncated concordance. Accordingly, in addition to offering a heavier information load and a lower item-difficulty level, his format seems to also reflect neither the commonly perceived notion of a concordance, nor hence the usual form of presentation of classroom concordances. In sum, if everything else is equal, then Butler's item prompt is likely to offer several more words per question than ConCloze is, and does not seem to conform to the usual sense of a concordance. This also means that, for example, the unique

advantages of concordances may not be delivered efficiently or may not be delivered altogether in Butler's items (cf., e.g., pages 6, 10 (below), and 12 for some benefits of concordances in pedagogical contexts).

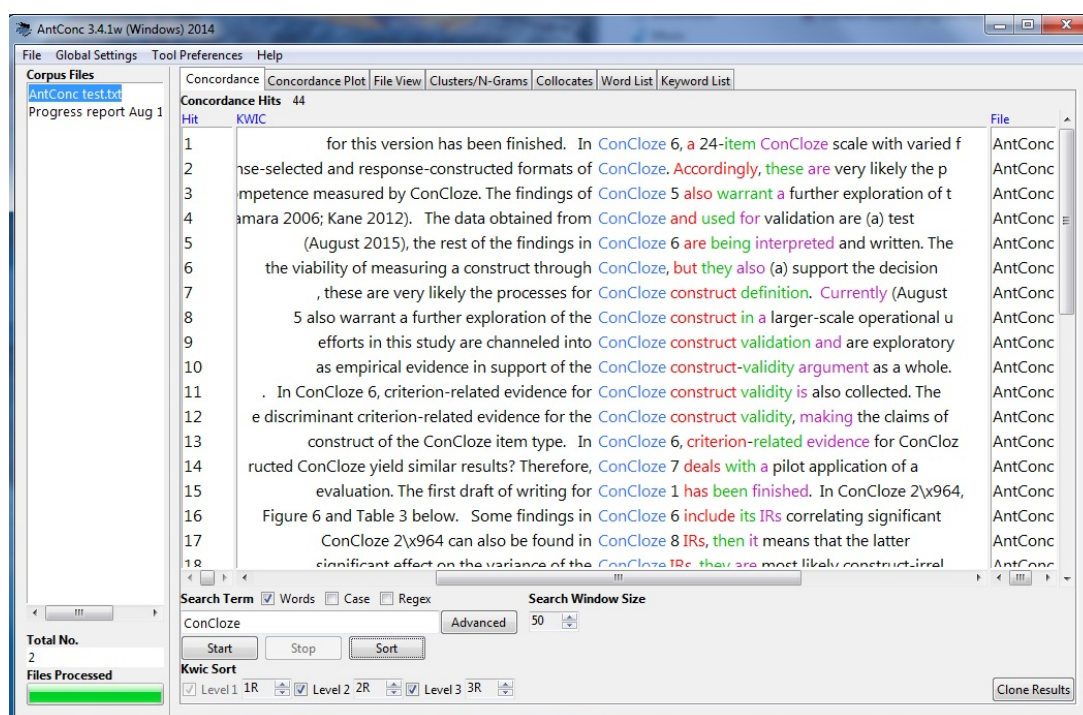


Figure 6 AntConc's (Anthony 2014) concordance view

Given the differing sentence alignments, a second distinction that potentially lies between the ConCloze item prompt and Butler's is about information processing. The point of interest starts with the gap position where the word being questioned is taken away. In ConCloze, the concordance prompt (as in Figure 1, page 1) features the central-vertically aligned gaps in the node position of all the concordance lines. This type of alignment may be conveniently called KWIC-centered, which is the typical display in concordancing (cf. also Figure 6 above for an example default-generated by a concordancer). The display draws attention to the KWIC position and allows a recursive reading of the words around the KWIC, facilitating an analysis of their linguistic structure and inductive observations (Papp 2007; Aull 2015). The display also makes patterns noticeable, disambiguates confusing senses, if any, and is thus easy and time-efficient for language analysis when read (Schmitt 2000; Barlow 2004; Gilquin & Granger 2010). By contrast, Butler's texts of the prompt (as in Figure 5, page 8) are all left-aligned, and each gap is distributed in the original occurrence of the word being tested on. An implication

is that the accessibility of Butler's prototype may, at the very least, be affected by the alignment, which is unconventional for a concordance. At worst, the aforementioned benefits of the concordances cannot be delivered as intended.

In light of the differing alignments and a consequent inability of Butler's (1991) item format to deliver concordance usefulness, patterns of information processing between Butler's and ConCloze's might be inherently different. In validity terms, an inference could be that there may exist discrepancy, rather than likeness, in terms of the domain of competence that is tested in the two tests. In sum, Butler's format of the concordance prompt is potentially distinct from ConCloze's in terms of both the information presented to the examinees and their patterns of information processing. Collectively, it may be argued that the two tests could have distinguishable constructs.

Butler (1991: 29) regarded his test as an "undeveloped technique." However, due to unspecified technical difficulties, Butler reported not proceeding with his testing program to another phase such as field-testing, and he did not make any construct interpretations either. On this account, a summary in favor of the ConCloze originality could be as follows. First, Butler's study is a prototyping one whereas this study is construct-oriented. Secondly, his prompt format is left-aligned while ConCloze's is traditionally KWIC-centered. Lastly, his prompt is presented non-truncated whereas ConCloze's prompt is conventionally truncated. All the distinctions drawn so far can suggest that the validity of score interpretations of the ConCloze format as it stands has never been investigated for a large-scale systematic testing. Accordingly, this study is likely providing an original account of its validity investigation.

The next recurrence of concordance-based cloze testing in the literature is Hargreaves (2000), who also worked for UCLES. Hargreaves highlighted the importance of vocabulary knowledge towards assessing language proficiency, and contended that depth of vocabulary knowledge can help to discriminate learners of different proficiency levels. A test item was offered in his discussion, which is illustrated in Figure 7 below (answer: *A remember*). The potential of such an item was claimed to be requiring the learners to show "greater knowledge of a word's properties and patterns" and assessing "dependent grammar patterns" as part of vocabulary knowledge (ibid.: 210f.).

Only one of the words in A, B, C, D is appropriate in **all three** of the blanks in the three sentences below. Please circle the appropriate letter:

A remember **B** agree **C** suggest **D** admit

She did not posting the letter.

She did not to post the letter.

She did not that she had posted the letter.

Figure 7 Hargreaves's (2000: 211) concordance-based cloze item

In Figure 7, the blanks are all aligned vertically but not centrally, meaning that Hargreaves's concordance is more similar to ConCloze (as in Figure 1, page 1) than Butler's (1991) is. Further, unlike Butler's constructed-response format, Hargreaves's item is a selected-response one, which is also identical to the format used in this study. However, in Hargreaves's, the three sentences of the prompt share a very similar message and differ primarily only in terms of verb valences. These prompt sentences, accordingly, seem somewhat unnatural as they may not be taken from naturally occurring texts. Given the questionability of the source of sentences, an inference is that Hargreaves's item prompt may be purposely written for testing, rather than corpus-derived, and may not reflect how the language is used in reality.

It is worth stating that authenticity is a key advantage which corpus-retrieved texts give to learners (e.g., Römer 2011; Flowerdew 2012). For example, concordances of authentic texts can raise awareness of lexicogrammatical patterns more efficiently than traditional deductive methods could (summarized in Coxhead 2010). Moreover, texts from corpora are significantly richer than simplified texts in terms of word frequency (Crossley *et al.* 2007). Considering such advantage, it can be argued that Hargreaves's task content does not realize the benefits which typical concordances have to offer (cf. McCarthy & Carter 1997; O'Dell 1997 for more benefits of authentic texts). On the one hand, authenticity does not automatically guarantee nor disqualify validity. Yet, it seems reasonable to assume that Hargreaves's prompt may be qualitatively different from ConCloze's prompt. In light of this potential difference, a possibility is that if the number of words were identical in the two prompts, their content and the ways examinees process it could yet be different. This would mean that their underlying domains of competence tested may not be identical.

Apart from the advantageous authenticity, the other distinction between Hargreaves's item prompt and ConCloze's is truncation. Hargreaves's sentences are not truncated, a manipulation of test-task content similar to Butler's. This indicates that Hargreaves's prompt format differs from the conventional concordance view and also from ConCloze. In validity terms, again, the sets of information offered to the examinees are likely dissimilar in the two cases and, therefore, their constructs may be significantly different from each other.

Hargreaves (2000) did not provide any specific validity evidence for the item format. Nor did he refer to specific research publications on his test item. Therefore, in addition to helping to establish a substantive distinction for ConCloze, reviewing Hargreaves (2000) also indicates that previous validation studies into the item type are unlikely to exist. This thus confirms the originality argument for the current validity investigation.

The last work pertaining to concordance-based cloze testing is a previous use of an item format by CELA (presently an UCLES department for English as a Second/Other Language Assessment (ESOL)) (2010a). Having five items, the format was in a section of Cambridge English: Advanced (also CAE), the sample item of which is displayed in Figure 8 below. Each of the items is found to consist of three complete sentences requiring the same word to fill out their gaps. Also, the sentences are left-aligned just as ordinary texts are. On the one hand, CELA did not explicitly name the item format as a concordance-based cloze. Yet, both Butler (1991) and Hargreaves (2000) referred similarly to their corresponding item formats in the context of UCLES, implying that the format by CELA is also likely to be a legacy of Butler's prototyping. Based on this interpretation about CELA's (2010a) item format, at least three distinctions in form can be drawn in contrast to ConCloze: truncation (non-truncated vs. traditionally truncated, respectively), alignment (left-aligned vs. KWIC-centered), and type of expected response (constructed-response vs. selected-response). Because CELA's (2010a) item format is similar to that by Butler's (1991) (cf. the earlier arguments, pages 9 and 10), the format is also likely to differ from ConCloze's in terms of the information processed and the pattern of information processing. This means that in addition to the differences in form, CELA's format may also invoke different substantive processing.

For questions **38 – 42**, think of **one** word only which can be used appropriately in all three sentences. Here is an example (**0**).

Example:

- 0** They say the new minister is a lovely person and very to talk to.
 My neighbours have not had a very life, but they always seem cheerful.
 It's enough to see why the town is popular with tourists.

Example:

0	E	A	S	Y												
---	---	---	---	---	--	--	--	--	--	--	--	--	--	--	--	--

Figure 8 CELA's (2010b) sample item in CAE Use of English

In light of the differences laid out thus far, it may be said that the item formats from Butler (1991) to Hargreaves (2000) and CELA (2010a) are not properly designated as concordance-based testing. Rather, the formats merely resemble gap-filling tasks featuring multiple prompts (cf. also Kongsuwanakul 2014b for discussion on the differences in task formats). Because CELA (2010a) does not provide references of related validation studies, it could also be inferred that there seem to have been no previous empirical studies into the validity of score interpretations for a concordance-based cloze test. On this account, an argument is that the ConCloze format is likely to be innovative.

Suggested in the rationale (page 1), a corpus-based test innovation should have potential to meet a need which has not been met. To date, there have been no studies that systematically turn concordance lines into test-task content in spite of a growing corpus use in linguistic analyses and language assessment (Kongsuwanakul 2014b; 2015b) (cf. also Part 2.2, pages 31ff. for a gap in the literature about roles of concordances in language assessment). Therefore, not only is the item type innovative, but there is also an original need calling for a research enterprise like the present study to systematically pioneer the design.

In addition to the originality of this study in responding to a need for corpus-based test innovation, the need is also supported by three additional reasons pertaining to language pedagogy. First, corpora and concordancing have a tendency for increased use in areas other than corpus linguistics, e.g., language teaching and learning (McCarthy & O'Keeffe 2010; Willis 2011). Accordingly, using concordances in language testing will be a natural extension to those areas to reflect this tendency. Secondly, concordances

composed of authentic texts have been mentioned to have several distinct advantages over texts aligned normally that are written specifically for testing (pages 10 and 12). For this reason, using concordances as test-task content may potentially mean an unprecedented delivery of their usefulness in the contexts of language testing—a theoretical projection which only empirical evidence could test.

The last reason pertaining to language pedagogy that supports an original need for a corpus-based test innovation is about utilizing concordances. ConCloze is made of corpus-derived texts, whose accessibility is increasing rapidly due to internet growth (Alderson 2000; Kongsuwannakul 2014a). Therefore, the item type may be deemed a new and timely option for non-native teachers of English and test writers, who are in a large number across the globe. In using concordances, they could benefit from becoming less dependent on native intuition in test writing (cf. Barker 2006: 3 for an increasing role of corpus-informed insights in place of native intuition in testing practices). In fact, the CAE-related test formats from Butler (1991) to Hargreaves (2000) and CELA (2010a) do not seem to properly use concordance lines as they stand in constructing item prompts. Accordingly, developing ConCloze items in this study would mean an initiative to make direct use of concordances in their accurate form as a main part of test-task content. In sum, the present study of ConCloze validity could address a need for corpus-based test innovation, reflect the tendency for increased corpus use, originally deliver usefulness of concordances in language testing, and serve as a timely option in test writing, which can particularly benefit non-native English practitioners.

In conclusion, concordances have been argued, *inter alia*, to provide richer word-frequency information than ordinary texts, allow recursive reading with inductive observations, and facilitate time-efficient analyses of lexicogrammatical patterns. Accordingly, the current study could be an improvement over existing projects of concordance-based tests in that this study seeks to originally use concordances in their accurate form as test-task content. This means that the item type may be able to deliver advantages of concordances in the testing contexts unprecedentedly, and address a need for concordance-based testing. How defining the score meaning of ConCloze would be an improvement over existing concordance-based testing projects in terms of validity will be discussed in Section 1.3.2.

1.3 Research Questions and Hypothetical Construct

1.3.1 Research questions

This research focuses on investigating the validity of score interpretations for the ConCloze item type. Given this, the main research question is: *What is the construct of ConCloze scores?*, which can be defined as: *What language process and knowledge does ConCloze tap into?* and *How do the examinees engage in ConCloze test tasks?* In no particular order, these questions are operationalized in Table 1 and explanations will follow thereafter.

Table 1 Operational research questions

No.	Question	Aspect
1	Are item responses internally consistent?	Structural
2	What domain(s) do the task engagements involve?	Content
3	What item component(s) do the examinees use in task engagement?	Substantive
4	What process(es) do the task engagements involve?	Substantive
5	Is item difficulty affected by variation in task content?	Substantive
6	Is there consistency in item responses and processes across occasions?	Generalizability
7	Are ConCloze scores significantly associated with Read's (1993; 1998) Word Associates Format (WAF) scores?	External

A premise at the beginning of this study is an unknown score meaning of the item type. Accordingly, the validity inquiry is largely exploratory: seeking to define the construct from unknownness. Investigating the ConCloze construct in this manner can be useful for two reasons. First, no studies are found to have explored a score meaning for this item type in its truly concordance-based form (cf. Part 1.2). Hence, allowing its item responses to inform the construct definition seems to be a careful course of action. Rather than setting a fixed construct only to confirm or reject via a statistical analysis of variance, the score meaning can be defined flexibly depending on how the tasks are actually engaged with (cf. Kane 2012b: 4 for an argument against a confirmationist bias in developing an interpretation to propose). Secondly, in the literature on language learning, domains of competence for which concordances are utilized can be found assumed substantively. Examples include knowledge of noun phrases modifying adverbials (Kongsuwannakul 2013), knowledge and use of English synonyms (Yeh *et al.* 2007), transferring academic-word knowledge to a writing task (Kaur & Hegelheimer 2005), interpreting a multifunctional phrase in a foreign language (Kenning 2000), familiarizing

oneself with lexicogrammatical contexts of academic English vocabulary (Thurstun & Candlin 1998), spelling words and choosing words for new texts (Cobb 1997), “extrapolating holistically from fragmentary evidence” (Stevens 1991b: 48), and analyzing learner errors post hoc (Butler 1990). Yet, nearly all of them seem to take face validity of the domains involved, rather than defining the domains afresh through verbalizations, for example. This implies that it would be very difficult for the present research to come up with a single and accurate hypothesis merely by extrapolating a score meaning that is obtained from a review of the concordance-related literature. Therefore, a pragmatic solution to the difficulty is set to be exploring the ConCloze construct afresh and enriching it with insights from previous studies where possible. The limitations of this approach will be discussed in Part 5.2 (pages 276ff.).

In defining a ConCloze score meaning, a fundamental assumption is that the item responses must contain interpretable patterns. In language testing, such patterns are considered observable realizations of an underlying competence at work (cf. Messick 1989 for roles of consistency and generalizability in test-score interpretations). Summarized in Table 1 above, the realizations can be categorized into five following aspects according to Messick (1995).

The first operational aspect is structural, inquiring if ConCloze items elicit responses that are consistent. For example, an examinee who has a high level of the target competence tested should likely be able to score well from beginning to end of the test. Such consistency in test performance can show that the responses are systematically structured, reflecting the measurement of the competence at issue. Question 1 in Table 1 is designed to seek validity evidence on the scale level as well as across other facets of construct measurement.

Question 2 in Table 1 deals with the content aspect of validity. This refers to the domain(s) and its boundaries and nature being defined as the construct. For example, proficiency in professional and academic English grammatical and vocabulary use is going to be hypothesized in the next section 1.3.2 (pages 22ff.) as an initial score meaning, which is the content domain whose nature can be scrutinized—rejected, refined, or attested—upon obtaining actual item responses. The aim of investigating this aspect is to (a) determine the area(s) of competence involved during task engagement, and (b) investigate the way these areas, if applicable, are tapped into to produce item responses.

Question 2 is designed to look for particular domains of language competence expressed in the content of verbalizations during task engagement as well as in the patterns of item responses.

Questions 3–5 in Table 1 address the substantive aspect of construct validity. This refers to the domain processes mobilized in test-task engagement. For example, if the aforementioned hypothetical domain *use of grammar and vocabulary* is invoked, then prospective processes could be those of selecting the words in the options and explaining their meaning. The substantive aspect also entails variations in responses that are theoretically expected, oftentimes as a result of manipulating task content. For example, adding more concordance lines could be expected to give more clues to the examinees. If analyzing item responses really indicates so, then an increase in scores caused by the addition is substantive-validity evidence. Question 3 is designed to determine the component(s) of the items that are systematically used for task engagement. This could then help to set the construct-relevant boundary of task content for the previous content-validity aspect. Question 4 is designed to create processes that can account for all or most of the verbalizations during task engagement. Question 5 is designed to model how variations in test-task content (e.g., more concordance lines, fewer options) would affect item difficulty. This model can then represent the processing that systematically determines the variations for task completion.

Question 6 in Table 1 deals with the generalizability aspect of construct validity, which refers to the extent of consistency with which item responses and processes are applicable across occasions. For example, when one test is administered on two different groups of examinees, and the two sets of scores correlate highly, it is likely that their scores reflect the same construct at work. Question 6 is designed to gather evidence in support of the claims for a power of generalization of the test-score interpretations to the universe of admissible scores and for test-task representativeness.

Question 7 in Table 1 deals with the external aspect of construct validity, which refers to the extent that the scores representing the construct domain are associated with the scores from another test. This is often as predicted by the theoretical framework of the main measure, which is ConCloze in this study. For example, when two tests are theorized to have parts of their competence domain in common, their scores should be found correlating positively. Question 7 is designed to investigate the variability of scores

as predicted by a model of the hypothetical construct of ConCloze. This model together with the rationale for selecting a criterion test will be explained in Part 2.3 of Chapter 2.

Given that the external-validity question has been discussed, it is also worth stating that the term *criterion test* has little to do with *criterion-referenced assessment*. The former term is often used interchangeably with *criterion-related test*, referring to an external measure that is brought into a particular research context for validity investigation. Usually already validated, the measure functions as another source of empirical inferences for interpreting item responses of the measure being explored. For example, the validity of ConCloze-score interpretations is investigated in this study, and Read's (1993; 1998) Word Associates Format (WAF) is brought in for exploring additional dimensions related to the ConCloze construct (to be detailed in Part 2.3, pages 47ff.). By contrast, *criterion-referenced assessment* is a paradigm for evaluating test scores, often by setting a particular standard or standards (cf. Henning 1987: 6ff.). For example, the examinees who score over 80% of the maximum possible score will attain an A whereas those below 50% get a C for their grades. This assessment paradigm is not used much in this study, and the details about the assessment paradigm used will be provided in Part 2.3 of Chapter 2.

1.3.2 Hypothetical construct

In Part 1.2 (page 6), it is argued that the ConCloze construct is likely to have never been investigated. In the absence of score meaning, a hypothetical construct will be formulated in this section by transferring the competence domain of previous concordance-based tests to the current testing situation. Determining a hypothetical score meaning can be useful for three reasons. First, when item responses are explored in Chapter 3 and Chapter 4, the meaning can enhance the extrapolation of score interpretations to the target domain of competence (cf. Chapelle 2012: 23f. for roles of a theoretical construct in the validity argument). Secondly, the hypothetical construct can help with forming a test purpose, which will be discussed later in this section. This implies that the fidelity between the hypothetical meaning and the outcomes of interpreting item responses can be systematically appraised in light of the test purpose, thereby refining the construct throughout the validation process. Lastly, as the ConCloze item type has been argued in Part 1.2 to be distinct from the other concordance-based cloze tests, the current

discussion can help to pinpoint how the present study would be an improvement to those existing tests.

The concordance-based cloze tests previously discussed are all related to CELA's CAE program at UCLES (page 14). Given this, the CAE construct as advertised on its website is worth considering because it may contain certain clues that could serve as a starting point for formulating the current hypothetical construct. Illustrated in Figure 9 below, CELA describes the construct of the test program as being able to communicate effectively in the English language for professional and academic purposes. It may thus be contended that the CAE score meaning is to reflect proficiency in professional and academic English use. For discussion purposes, it is also worth differentiating two terms referred to frequently: *competence* and *proficiency*. The word *competence* is used in this study as a generic term for an area or domain of language ability or knowledge that is internalized through learning and acquisition (cf. also Bachman & Palmer 2010: 33 for *language ability* as a superordinate term for language users' communicative capability). In comparison, when a particular domain or a combination of domains is mobilized for task completion, the test performance, particularly that on the scale level, may be interpreted as reflecting the respondents' *proficiency* levels in using the domain, as in 'proficient language users' vs. 'inept users' (cf. also Cummins 2000 for an extensive discussion of *language proficiency* as an intervening variable between social interaction and contextualized learning, and academic performance). Measuring the construct proficiency will be conceptualized in Part 2.3 of Chapter 2.

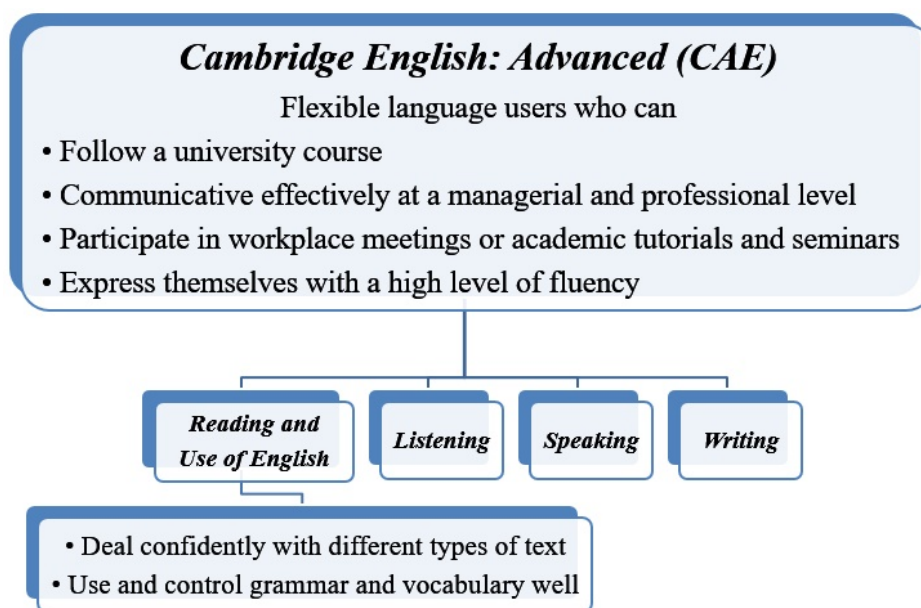


Figure 9 CELA's (2016a; 2016b, adapted) 'can-do' construct of the CAE program

CELA's (2010b) concordance-based cloze test was incorporated as one of eight test formats in a previous version of the CAE Reading and Use of English section (cf. also pages 14f. for the concordance-based format). Shown in Figure 9 above, CELA (2016a; 2016b) gives a collective meaning for the section scores as being able to deal with different types of text with a good command of grammar and vocabulary. However, CELA does not provide a separate meaning for each of the formats in the containing section. Nor does it specify (a) if each format can individually serve as a psychometric instrument measuring the same construct, or (b) if the construct is composite and each format measures a distinct part of it. Because different test formats may invoke different aspects of a competence domain or different domains altogether, criticism could be that the meaning given by CELA is quite generic and may not apply equally to all of the formats used. Given this lack of specificity, it would be deemed vague if ConCloze adopts the CAE Reading and Use of English score meaning as it stands. An argument is that the ability to fluently command English grammar and vocabulary cannot be taken for granted as an ideal candidate for the ConCloze construct.

In addition to the vagueness, directly adopting the score meaning from the CAE test program could also be psychometrically problematic. The construct of its Reading and Use of English section involves grammar and vocabulary—language domains known to be dimensional (e.g., Purpura 2004; Hancioglu et al. 2008; Alderson & Kremmel

2013). Taking into consideration the non-specificity described above, it seems that the construct could be so multidimensional in its domains (e.g., collocation, connectives) that the validity investigation may turn out to be highly complicated and unmanageable. Accordingly, proficiency in professional and academic English grammatical and vocabulary use would be inappropriately broad as a construct of a particular test format like ConCloze, and can only serve as a superordinate meaning for validation purposes. In other words, while the score meaning derived from the CAE program may be able to encompass the ConCloze construct, it could be due to the fact that it is very broad. This means that the construct related to previous concordance-based tests needs clarification, insofar as the meaning applies specifically to concordance-based cloze tests, rather than additionally to several other test formats.

In light of the meaning obtained from the previous CAE-related tests, two major points can be made for this validity investigation. First, a hypothetical construct for the ConCloze item type could be proposed as proficiency in professional and academic English grammatical and vocabulary use. This construct is not intended as a finite representation, but rather as a starting point that allows appraisal and refinement in the rest of this thesis. Secondly, the previous concordance-based cloze tests are not provided with a validated score meaning (page 14), and the CAE program only provides a superordinate meaning for their containing section (Reading and Use of English). Accordingly, it may also be argued that this study could be an improvement to the lack of specific score meaning in those tests. In fact, it is worth reiterating that a corpus-based test innovation should have the potential to measure a competence that has never been measured before (page 1). Because CELA does not elaborate the construct of each test format used in the CAE program—particularly separate ones for those Reading and Use of English formats, which include a concordance-based test (page 14)—this research project may thus be considered an improvement to the test program. Namely, the ConCloze item type has the potential to measure competence domains that have never been specified for concordance-based testing before.

Given that the hypothetical construct is set initially as proficiency in professional and academic English grammatical and vocabulary use, an immediate purpose of ConCloze would be to assess the proficiency. Administering a language test in order to learn if a student has a high or low level of an intended ability and hence their proficiency in a particular domain means that the test is being used for diagnostic purposes (Henning

1987: 1–2). That said, it is worth pointing out that the diagnosis mentioned is essentially to measure the examinees' performance and place their scores along a ConCloze scale for proficiency estimation, yet not to identify their problematic area(s) of the intended construct. This is so because construct definition is the primary research focus and will have to take place before construct-relevant subdomains can ever be identified. Accordingly, the purpose of ConCloze can be expressed initially as a measure to assess and give feedback on examinees' proficiency in professional and academic English grammatical and vocabulary use. The framework for assessing the language-use proficiency through ConCloze will be set in Part 2.3 (pages 33ff.) and then supported by a psychological mechanism in Part 2.4 of Chapter 2 (pages 55f.).

In light of the main test purpose, another line of inquiry also deserves clarification. The literature-informed construct has been formulated and is subject to refinement by appraising its fidelity with results of analyses of actual item responses (page 19). This means that assessing the proficiency would take a leading role in the construct appraisal because the validation directly relies on it. By contrast, reporting the scores to the participants does not directly affect the interpretations of item responses but merely involves test logistics. For this reason, the aforementioned purpose of giving score feedback to the test takers could only have a secondary role in the appraisal.

In addition to the main purpose as a proficiency assessment tool, setting the hypothetical ConCloze construct also delimits the scope of the test purpose. Developing an argument which justifies score interpretations for the item type is a priority in this study (pages 1f.). Hence, the test format is decided not to be tied in to one particular curriculum or pedagogical program. Validating the score meaning without being closely related to a specific curricular component would mean a relatively neutral test purpose that reduces sociopolitical influences in the test framework (cf. Chapelle 2012: 24f. for (a) intended test decisions, (b) indirect, multi-directional score interpretations, and (c) standard referencing as common controversies caused by emphasizing test use over test interpretation). As such, the power of generalization of the construct that is going to be defined in this study will be greater than a construct with links specific to one particular test program (cf. Messick 1989; Kane *et al.* 1999 for trading off between generalizability and extrapolability of score interpretations in validity arguments). Accordingly, this study focuses more on the interpretive arguments than on test-use arguments, a framework whose limitations will be discussed in Part 5.2 (pages 275ff.).

In summary, the ConCloze item type has been argued in Part 1.2 to be an improvement over other concordance-based testing projects in that it seeks to use concordances in their accurate form and potentially be able to deliver their usefulness in the testing contexts. In this part, the current study is argued to be an improvement over those projects in that it is going to define the score meaning that is specific to concordance-based testing. In investigating the validity, the purpose of the test is initially set to be a proficiency test measuring professional and academic English grammatical and vocabulary use. This construct is subject to refinement throughout this thesis.

1.4 Potential Impact

Literature on test writing, particularly concerning pre-operational testing, is an “underappreciated child of the test development process” (Kenyon & MacGregor 2012: 305; cf. also Read 2012: 308 for a similar idea). Moreover, test- and item type-development processes are usually kept confidential in the testing industry. Given this, the process described in this study (Chapter 3 to ConCloze 6) can be useful for other test-writing programs because its test specifications (specs) and the results of task-content modifications are supplied in detail. This means that course developers and teachers alike can adjust the blueprints for writing a new ConCloze which suits their needs. For example, a ConCloze whose prompts are based on concordance lines retrieved from course textbooks may be produced assessing the students’ use of technical language for the right contexts—an edition of ConCloze which could thus serve a diagnostic purpose for feedback on their learning achievement. Because there are a large number of native and non-native language practitioners across the globe, it may be stated that this study can have a huge impact by giving a building block to their ConCloze-writing practices (cf. also page 14 for ConCloze as an original item format to deliver concordance usefulness).

Apart from providing detailed blueprints and interpretive results for ConCloze remodelling, another related impact is on classroom assessment. For example, in Figure 10 below, the KWIC display in Figure 6 (page 10) is turned into a KWIC-blanked-out display (answer: *ConCloze*) using an available function called ‘Hide search term in KWIC display’ (Anthony 2014). Upon one click, the occurrences of the KWIC are all hidden, and picking suitable concordance lines from the display would be most of what is needed

in producing an item prompt (cf. Stevens 1991a: 37 and 40, respectively, who referred to this step as requiring ‘minimal effort’ and ‘mechanical’). With a validated score meaning, item distractors can be selected meaningfully for low-stakes testing such as in-class quizzes and formative assessment.

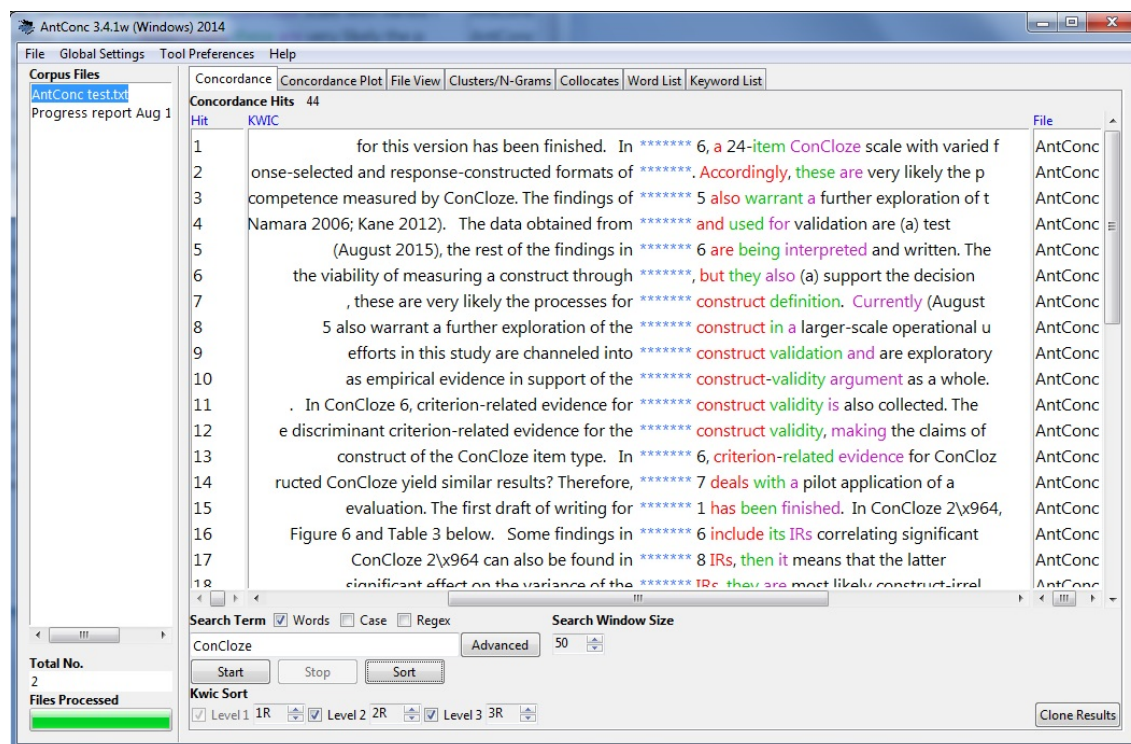


Figure 10 Hiding the KWIC in an AntConc concordance

The significance of the example given above lies with its impact on item generation by those who have hands-on experience with corpora and concordancing. A message is that their existing tools are useful not only for language teaching, but also for in-class assessment. If the course textbooks that the students need to read are already available electronically, a purpose-built corpus can be easily arranged (cf. Tribble & Jones 1990; Tribble 1997; Cobb 1997, cited in Nation 2001: 112 for concordance use in many language-learning activities but systematic testing). Being able to apply corpus-based tools to test writing can be important because test writers’ dependence on native intuition in test writing can be lessened with concordance use (cf. Belcher 2006 for discussion about roles of concordances in exploring use of English for specific purposes). Given that non-native English teachers currently outnumber native ones, such application

could be deemed particularly liberating; the teachers who routinely use corpora for pedagogical purposes would find writing a customized ConCloze test a natural extension of their corpus use (cf. also page 14 for ConCloze as an original item format to serve as a new option in test writing). In summary, impacts of this study include giving a building block for language practitioners to write tailored ConCloze items and offering a practical corpus-based tool for non-native teachers in classroom assessment.

1.5 Ethical Considerations

This chapter has thus far provided a prologue to the current validity investigation. Since this research involves human respondents and copyright material, for good practices, ethical issues will also be addressed here. The issues can be divided into three areas: anonymity and confidentiality of participants (1.5.1), use of incentive (1.5.2), and copyright material for test construction (1.5.3). This study has received ethical approval from the University of Leicester.

1.5.1 Anonymity and confidentiality

This research follows the British Educational Research Association (BERA)'s (2011) guideline on treatment of participants' data. This means that identities of the examinees remain protected throughout the thesis. For quantitative parts, this manifests itself as aggregate score reports and analyses; no item responses are identified as produced by particular examinees. For qualitative parts, pseudonyms are used in reference to the participants where necessary.

1.5.2 Incentive use

There are two main categories of incentive used in this study: physical and psychological. Concerning physical incentives, refreshment is offered to ConCloze 2–4 participants, and a remuneration to ConCloze 7 participants. Regarding psychological ones, the following are offered to all participants, the first of which is complete confidentiality and anonymity. Secondly, it gives words of compliment for their contributions to the test development and to acquisition of knowledge. Lastly, it sends

individual score reports directly to their inboxes within 14 days upon test completion (applicable only to ConCloze 1 and 5–6). According to BERA (2011), negative effects should not be implicated in incentive use; the incentives described here are not known to cause harm to the respondents' health and safety. Therefore, it can be contended that this study complies with the BERA's guideline on using incentives for encouraging research participation.

1.5.3 Copyright in concordance texts

The test items in the study contain texts retrieved from a corpus, which are naturally occurring and generally copyright-protected. However, the texts are used here for research purposes, not commercially, and in a snippet-like way. On this account, the use can be deemed a fair use according to UK copyright law (cf. UK Copyright Service 2004; 2009 for details of the law). Hence, it is unlikely that an infringement of copyright material is incurred in this study.

1.6 Concluding Remarks and Thesis Structure

In this chapter, the context of this study has been provided. It starts with the cruciality of investigating the validity of score interpretations for the ConCloze item type. The test format as developed in this study is argued to be innovative, and an original account of its validity investigation is being provided. At the beginning of this thesis, the test construct has been set initially as professional and academic English grammatical and vocabulary use, and the test purpose as a proficiency test, focusing primarily on the assessment of the construct domain. In the rest of this study, the construct will be defined afresh from item responses and theoretical grounds, so as to refine the test purpose which is specific to the concordance-based cloze testing.

In light of the innovativeness, investigating the score meaning of the test format could have an impact on both testing research and practices. To begin with, the previous concordance-based tests are not supplied with a validated and specific score meaning, and so specifying the score meaning of ConCloze could address this issue of a lack of domain specificity. Unlike the previous tests, this study makes use of concordance lines in their accurate form as part of task content and therefore may deliver the advantages of

concordances in testing contexts unprecedentedly. For example, non-native teachers of English would find online corpora and their concordancers particularly useful because they can provide them with concordance lines that can be used for item generation. This means that the domain-specific meaning of the ConCloze scores to obtain in this study could be deemed an improvement to the previous tests and may be able to apply widely to other newly developed concordance-based cloze tests, a generalizability aspect of construct validity.

The rest of this thesis is structured as follows. In Chapter 2, a gap will be identified in a literature review on language assessment. A theoretical conception will then be laid out for the item type: a framework for construct measurement and in relation to a criterion test (Read's (1993; 1998) Word Associates Format (WAF, aka. Word Associates Test)), and the hypothetical competence underlying the ConCloze performance. Afterwards, the test development process will be reported in Chapter 3, which begins with a quantitative prototyping (ConCloze 1), where initial evidence of the underlying competence is sought. It is hypothesized that the item responses have internal consistency, seeking to initially answer structural-validity Question 1 of Table 1 (page 16). Then a qualitative prototyping (ConCloze 2–4) follows: verbal reports will be collected, and task engagement observed and analyzed. It is hypothesized that (a) the item prompt and options are used in test-task engagement, and (b) there is a core domain and processes underlying task engagement. These interpretations will answer content- and substantive-validity Questions 2–4 of Table 1. Finally, the field-test (ConCloze 5) seeks quantitative evidence from a large sample of respondents, seeking to answer substantive-validity Question 5 of Table 1. It is hypothesized that varying test-task content changes item difficulty.

In Chapter 4, responses to the operational test (ConCloze 6) will be analyzed, beginning with test equating and concurrent item calibration. Answering generalizability-validity Question 6 of Table 1, a hypothesis is that the ConCloze 5–6 item responses can be put in the same model. Further, the criterion test is also administered, seeking to answer external-validity Question 7 of Table 1; ConCloze scores are hypothesized to be significantly associated with WAF scores. Afterwards, more qualitative evidence will be sought in test applications (ConCloze 7), seeking to fine-tune substantive-validity interpretations and deal with the generalizability aspect of Table 1. These include (a) applying the processes constructed to new items, (b) contrasting high–low ConCloze

performers with face-validity evidence, and (c) applying the original item format to a different test mode and type of expected response.

Finally, this thesis concludes in Chapter 5 with a summary of what is learned from empirical evidence about ConCloze engagement. A construct model will also be generated in light of the theoretical framework from Chapter 2. Also, whether the gap identified in Chapter 2 has been filled out and whether the test-score interpretations are adequate and appropriate will also be appraised. Chapter 5 then ends with recommendations for future research.

Chapter 2 Literature Review

2.1 Introduction

Chapter 1 has set the background for this study by showing the originality of the ConCloze item type. To date, the concordance in its typical presentation—truncated concordance lines and a centered KWIC—has never been systematically used as test-task content. Nor is any research project known to investigate the validity of ConCloze-score interpretations based on a large-scale test administration (other possible names included). In the absence of direct literature, use of concordances in language assessment will be reviewed first in Part 2.2. When considered together with the originality argument in Part 1.2 (pages 4ff.), a significant gap that this study is intended to fill out will be identified, and where this study would be in the literature on a broader scale will be located.

In validity investigation, administering test versions to the respondents needs a framework for construct measurement. Accordingly, in addition to identifying a gap in the literature, a validation framework will also be laid out in this chapter. Taking into consideration both the hypothetical construct and the test purpose (Section 1.3.2, pages 19ff.), Part 2.3 conceptualizes linguistic qualities in the concordance-based prompt and how they are related to and useful for an efficient measurement of the hypothetical construct. Because of the concordance advantages, ConCloze will be proposed as an improvement over usual cloze tests and traditional language-test formats. Lastly, a primary assessment paradigm together with a criterion measure will also be set out in this part.

In ConCloze, the item prompt features a column of gaps in the KWIC position (cf., e.g., Figure 1, page 1). This can be deemed a feature derived from cloze testing—the words being questioned are *clozed*. Accordingly, in addition to laying out a theoretical framework for construct measurement, a procedure underlying cloze testing will also be reviewed in this chapter (Part 2.4). The aim is to argue for a new context of application of the cloze procedure to language testing. Moreover, lexical priming, a recent theory in language acquisition and use, will also be proposed as a genesis of the cloze procedure. This is in order to account for the operability of ConCloze and how ConCloze examinees can engage in the test tasks. Applying the theory of lexical priming to

ConCloze can be useful because it makes a new case for the theory in the settings of language testing.

2.2 Concordance Use in Language Assessment

In establishing the originality of this study, previous concordance-based cloze tests have been discussed in Chapter 1 (pages 8ff.). A broader literature about functions of concordances in language assessment will be reviewed in this part. The aim is to identify a gap in relation to the roles of concordances in language testing. The review can be divided into two categories: UCLES-related and off-UCLES. With regard to the UCLES-related literature, three previous concordance-based cloze tests have been discussed in Part 1.2: Butler (1991), Hargreaves (2000), and CELA (2010b), which are all involved with UCLES. The present review will proceed from those origins and focus on all the Cambridge-ESOL research notes since 2000 to date (61 issues altogether, available at <http://www.cambridgeenglish.org/research-and-validation/published-research/research-notes/>). The reason for focusing on the research notes at UCLES is that if concordances are used for cloze testing or in other test formats, UCLES could be the place from which references could most likely be tracked down. Table 2 shows the results of this literature survey.

Table 2 Works at UCLES related to concordancing and concordances

Work	Concordance-related Content
Porter-Szucs & Jameel (2014)	AntConc concordancer in identifying nativelike formulaic <i>n</i> -grams in a corpus of role-play transcripts
Boyd (2011)	Concordances as intervention strategy in vocabulary activities for acquisition
Capel (2010)	Concordances for investigating the frequencies of individual senses of words in the Cambridge International Corpus
Elliott (2010)	Concordances for investigating noun phrases and their genre tendency and associated semantic prosody
Proudfoot (2010)	Concordances for further exploring lexical-verb use in a corpus
Green (2008)	WordSmith concordances for comparing contexts of key words in each level of CEFR (Common European Framework of Reference)
Neff-van Aertselaer (2008)	WordSmith concordances for metadiscourse features (e.g., logical connectors) in English–Spanish corpora of expert writers
Rose (2008)	Concordances for in-depth studies of the key words in a frequency-based wordlist
Wright (2008)	Concordances for identifying shared collocations of the key words in three corpora
Barker (2006)	Concordancer as a tool in Cambridge Learner Corpus
Hughes (2006)	Concordances for comparing nominal groups in FCE (Cambridge English: First) reading texts with those in the British National Corpus (BNC)
Read (2005)	Concordancing as part of WordSmith Tools for corpus analyses
Hawkey & Roger (2001)	WordSmith concordances for examining language features in an IELTS (The International English Language Testing System)-examinee corpus
Boyle & Booth (2000)	Developing the Cambridge Learner Corpus, equipped with search & concordancing functions

From Table 2, concordances appear to have long been used at UCLES. They are utilized mainly for analyzing non-native speakers' language as well as comparing it with native speakers'. A synthesis of these research notes is that in testing contexts, concordances are generally recognized for their utility in investigating various linguistic features, usually as an extension to corpus-based inquiries. However, there is no reference or citation found concerning their applications as part of the test-task content in the same way as this study uses. This implies that validity investigations into concordance-based cloze tests, if existing, may not be known to the UCLES research team and affiliated scholars. Otherwise, there has been no validation program into the item type before. Accordingly, an argument is that the roles of concordances in the studies affiliated to UCLES are restricted to consultation, linguistic analysis, and language-learning enhancement. This thus indicates a gap in the literature on concordance-based cloze testing.

In addition to the publications affiliated to UCLES, surveying a wider scope of off-UCLES literature on concordances in language assessment yields a similar result. The discussion may begin chronologically with Alderson (1996), who suggested a number of ideas for applying concordances to language testing. For example, concordances can be presented for the examinees to make judgments about their genres, particular word classes, or meaning of certain text sequences. However, Alderson did not mention any use of concordances as part of cloze or gap-filling test-task content. Likewise, Schmitt (1999) pointed out that in the contexts of TOEFL (Test of English as a Foreign Language, a testing program by the Educational Testing Service located in America), corpora can be used, for example, to inform the test writers about words deserving testing and the most frequent meaning senses of particular words. Yet, Schmitt made no specific mention of concordances as part of test-task content. Also, Taylor & Barker (2008) only cited Hargreaves's (2000) suggestion about developing new corpus-based test formats, yet they did not provide any more detail about the formats (cf. Part 1.2, page 11 for discussion on Hargreaves's item). And recently, Park (2014) discussed use of concordances only as a tool informing course developers and teachers of various corpus-derived linguistic features. Concordances can also be used for analyzing learner-language aspects such as syntax, lexis, and cohesive devices (*ibid.*). Park did not, however, mention nor make reference to turning concordances into test tasks.

In summary, reviewing the literature on roles of concordances in language testing and assessment helps to ascertain the legitimacy of the originality claim in Chapter 1 (Part 1.2). Since Butler's (1991) prototyping of a concordance-based test, there seems to have been no recognition of the potential of concordances in forming test items. The lack is especially evident when reviewing in this chapter the functions of concordances in language assessment on a broader scale. Therefore, the following claims are very likely valid: there is a gap in the literature on applying concordances as test-task content, and the current research is making a unique addition to the literature on concordance use for language testing.

2.3 Framework for Construct Measurement

In Chapter 1, the purpose of ConCloze testing has been proposed as a proficiency test on professional and academic English grammatical and vocabulary use (page 23). In

this part, some elements of the test purpose will serve as a basis for explaining why ConCloze would be a receptive measure potentially suitable for this competence. To begin with, it has been argued that the hypothetical construct—derived in the form of a superordinate domain belonging to the CAE Reading and Use of English section (CELA 2010a)—is very broad and could be multicomponential in nature (cf. pages 21f. for an argument on the hypothetical construct; cf. Housen *et al.* 2012: 1; Dörnyei & Ryan 2015: 11 for a similar concept from the perspectives of second language acquisition and learner psychology). For example, categorized in Table 3 below, receptive proficiency in word use alone could involve several aspects of knowing a word, such as grammatical functions and collocations. An implication is thus that it seems to be impracticable to subjectively pick one or a few aspects of grammatical and vocabulary use and claim them as the construct domains specific to the concordance-based cloze testing. In fact, the notion of language proficiency is best expressed as concerted efforts of multiple domains of competence (Chapelle 2012). This means that multicomponentiality is inherent in the construct proficiency and could be anticipated in assessing it.

Table 3 Aspects of knowing a word receptively (Nation 2001: 27, adapted)

Aspect	Component	Question Asked
Form	Spoken	What does the word sound like?
	Written	What does the word look like?*
	Word parts	What parts are recognizable in this word?*
Meaning	Form and meaning	What meaning does this word form signal?*
	Concept and referents	What is included in the concept?*
	Associations	What other words does this make us think of?*
Use	Grammatical functions	In what patterns does the word occur?*
	Collocations	What words or types of words occur with this?*
	Constraints on use (register, frequency...)	Where, when, and how often would we expect to meet this word?*

* indicates Kongsuwannakul's (2014a) hypothesized inclusion in the construct of concordance-based cloze testing.

Apart from the multicomponentiality inherent in the construct proficiency, its origin seems to likewise call for a careful consideration of its components. Namely, the hypothetical construct relies on the score meaning of the containing section of the CAE program (Reading and Use of English) despite the fact that its test format that was claimed to be concordance-based is not identical to ConCloze (pages 14f.). Hence, it may be

considered precarious to presume for ConCloze a complete legitimacy in all of the program's multiple linguistic components without a proper construct investigation. In sum, (a) the multicomponential nature of the proficiency in grammatical and vocabulary use, and (b) the fact that the format of ConCloze differs substantially from that of the concordance-based format in the CAE program requires this validity study to seek the aspects of language use that are specific to ConCloze testing. Accordingly, an assumption in relation to the construct proficiency is that aspects of grammatical and vocabulary use specific to the ConCloze item type exist and could be identified with empirical item responses.

In addition to the multicomponentiality, the hypothetical construct also involves professional and academic English. The professional and academic genre consists of multiple discursive dimensions that are interconnected and virtually impossible to use in isolation (Bhatia 1993; 2008; Hancioglu *et al.* 2008). For example, as part of the professional dimensions, both sociocultural and psychological understandings may be intertwined and required when writing job-application letters. In those situations, recognizing the hierarchy of power needs to be expressed properly in a specific cultural context (Bhatia 1993: 118ff.). An argument is thus that real language use is so complicated that the language as it stands—rather than language written specifically for testing and overly simplified—should serve as a main part of task content when testing on language use. This means that for the purpose outlined earlier for ConCloze, seeking text types that embody multiple linguistic components and are embedded with interconnected dimensions of discursive patterns is of importance to a construct-representative assessment.

Given the mission of meeting the two prerequisites (multicomponentiality and discursive interconnectedness), concordances seem to be a most suitable choice for assessing the construct proficiency for two reasons. First, concordances are a platform for displaying texts which are retrieved from a corpus of authentic language and accordingly could reflect real language use (e.g., Stevens 1991a; Stubbs 2002: 62; Poole 2011: 2). In particular, concordance texts are widely recognized as containing rich information about the lexicogrammatical patterns of words (Nation 2001). While it seems impossible that concordances would be able to represent all language aspects in real use, lexicogrammatical patterning is a textual quality that is dynamically represented in authentic texts rather than in artificial ones (Oller 2005) (cf. also Part 1.2 for more

examples of concordance advantages). This means that concordances can serve the very purpose of the proficiency assessment because their naturally occurring texts are those that are likely to reflect multiple language components and interconnected discourse features. Moreover, in assessing language proficiency, task content that is context-embedded is generally recommended (Read 1993: 357). Because the ConCloze purpose is to serve as a proficiency test, it can thus be contended that concordance texts, which are naturally context-embedded, would be a promising option of a text type for the purpose. In sum, the hypothetical construct involves grammatical and vocabulary use, and so incorporating concordance lines into test-task content could at least in part fulfill the aforementioned prerequisites for construct representativeness. By reflecting lexicogrammatical patterns in real use, testing with concordance-based task content should allow an efficient measurement of the construct proficiency.

In addition to their authenticity, a second reason that concordances are a suitable choice to use as task content for assessing the construct proficiency lies in their format. Illustrated in Figure 1 (page 1) and Figure 3 (page 5), each concordance is composed of texts retrieved from multiple sources, a unique distinction that is not found in a usual cloze passage. This distinction could be significant for testing language use because incorporating texts, particularly short ones, by several writers is a key to neutralize idiosyncratic features and increase representativeness of the texts contained in a corpus (summarized in Coxhead 2000: 13). In Figure 11 below, this is when the concordance prompt serves as task content stimulating the examinees to engage in the task with a quasi-random sampling of authors' idiosyncrasies (details of sampling to be discussed in Chapter 3). In fact, it is also worth restating that cloze testing usually suffers a domino effect of having multiple blanks per passage, which is not known to apply to concordance-based cloze testing (page 6). Therefore, because a primary component of each ConCloze item is a concordance, the item type could be deemed to represent relatively neutralized features of language use without domino effects and so an improvement over existing cloze formats.

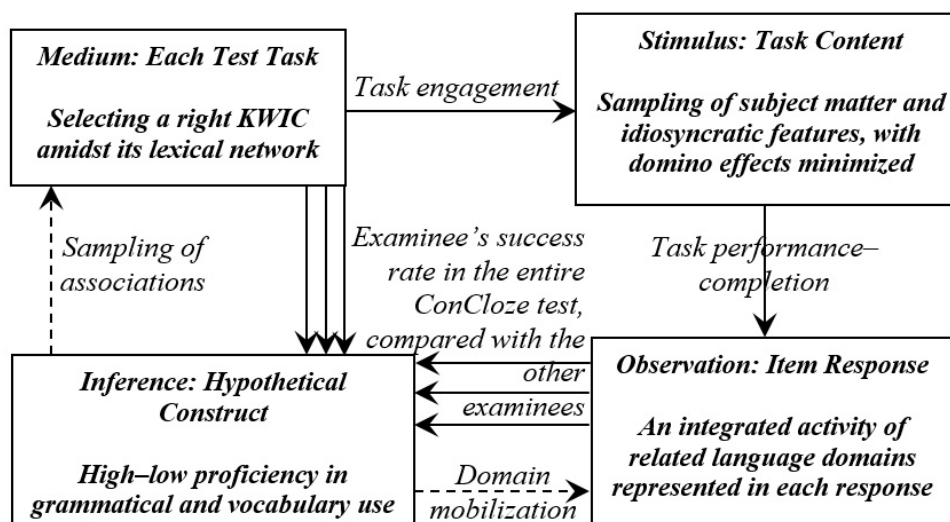


Figure 11 Measuring the construct proficiency with the ConCloze item type

With regard to task content of multiple sources, it is also worth pointing out that in real language use, the learners and language users may not necessarily need to always comprehend the texts in their entirety. An example is when they skim through academic papers for a piece of information. Each concordance-based prompt displays snippets of text from more sources than language test items in general, which usually have one or a few text sources per item. In Figure 11 above, this is when ConCloze-task content is presented with the subject matters sampled in the concordance lines, with idiosyncratic features balanced out. Therefore, processing a ConCloze task might involve going through multiple texts in a comparable fashion to reading different texts consecutively and thus might be able to usefully reflect the target domain of language use. Such processing can then determine implicitly if the examinees could pick up pieces of information, which at this stage of research are about grammatical and vocabulary use. For this reason, ConCloze might be argued as an improvement over existing language tests in that it may enable a relatively balanced sampling of multi-source task content – potentially a desirable characteristic in test design for its efficient processing and lower bias in subject matter (cf. also Part 1.2 for easy processing as a concordance benefit).

In summary, the hypothetical construct has been discussed thus far in terms of its linguistic qualities that concordances are likely to possess (i.e., multicomponentiality and discursive interconnectedness). Because the concordance is an essential component of ConCloze (page 1), four interrelated aspects of usefulness of those qualities have been

proposed as belonging to the ConCloze prompt: reflecting real language use, sampling texts in a comparatively unbiased manner, neutralizing authors' idiosyncratic features, and being better representative of language use. Without domino effects, the ConCloze item type is also argued to be an improvement over usual cloze formats. With a sampling of text sources that are presumably balanced out, it is also argued to be an improvement over traditional language-test formats. With those benefits, the item type is likely to be a suitable choice for assessing the construct proficiency in professional and academic grammatical and vocabulary use (cf. also Part 1.2 for an argument for ConCloze as an improvement over existing concordance-based test formats).

The ConCloze qualities and benefits described above are presumed retained in the testing contexts. Given this, some of them will be aligned to the purpose of the CAE program. Discussing how these qualities fit the program can be useful because it could help to ensure its usefulness on the industrial scale. In other words, the CAE-derived construct is used in this research as the initial score meaning which is subject to refinement upon interpreting actual item responses (Section 1.3.2). Yet, aligning the ConCloze qualities to the CAE framework would mean the item type belongs to a test model that is well-established in the testing industry.

There are two aspects of alignment to the CAE framework that are brought about by the ConCloze qualities. First, CAE is developed with a view to encouraging use of language skills in real-life situations (CELA 2010a). Using concordances of authentic texts in ConCloze would thus mean real-life situations conveyed in the task content, fitting the original intention of the CAE program. The second aspect of alignment is that CAE examinees "are expected to be able to understand texts taken from a range of sources" (CELA 2010a). The concordance prompt of ConCloze contains texts from multiple sources and thus could also serve this CAE objective. In sum, the construct framework of ConCloze is based on the fact that each ConCloze prompt contains a sample of multiple authentic texts, which is likely to enable the item type to fit the original framework of the CAE program (cf. Part 1.2 for most of the discussion on the benefits of the concordance and of the item type). On these accounts, the ConCloze item type seems to both fit the qualities of the target proficiency and the purpose of the CAE program.

In addition to the alignment of ConCloze qualities to the CAE framework, how the validity investigation is going to take shape is also worth conceptualizing. Laying out

fundamental concepts underlying the framework can be useful because it gives directions for research design and data analysis in later chapters. Given the newness of the item type, a first concept is that caution needs to be exercised in the validity investigation by producing multiple versions of the test, the process and rationale of which will be detailed in Chapter 3 (pages 66f. for the rationale). Using multiple test versions is, *inter alia*, in the interests of research resources, such that the measure can be prelaunched, and improved if necessary, with smaller samples of respondents before operational use (cf. Nissan & Schedl 2012 for steps of pre-operational testing).

The ConCloze purpose is to serve as a proficiency test (page 23). Given this, another concept underlying the validation framework is that the respondents' varying levels of the proficiency would be measured in test administration, insofar as differing total scores could be observed. Observing the respondents' differing scores is theoretically expected because it enables the interpretation that those with a high score are more proficient in the construct competence than those obtaining a low score, and vice versa (to be detailed later; cf. also page 20 for a distinction between *competence* and *proficiency* made in this study). Illustrated in Figure 11 (page 37), this is when the responses to the items in the entire test result collectively in varied total scores, thereby allowing an inference about their differences in the proficiency levels. Accordingly, a fundamental assumption for the operability of the current validity investigation is the commonality of the construct competence among ConCloze respondents (cf. Messick 1989 for psychological traits as common causes for test behaviors).

In light of the assumed commonality of the construct competence, an assessment paradigm can be selected meaningfully. There are two major categories of assessment paradigms usually referred to in the literature: norm-referencing and criterion-referencing. Norm-referencing is chosen as a main assessment paradigm in this study in two following dimensions, and its appropriateness will be discussed thereafter.

The first dimension in which norm-referencing is applied is the person view. Introduced earlier, the respondents are assumed to commonly mobilize the construct domain and related processes. Their competence mobilization in each task, as outlined in Figure 12 below, is expected to result in an item response: either scoring a point or failing to do so. An examinee's total score (i.e., the total number of test tasks the examinee has managed to deal successfully with) ranges from high to low and can be compared with

the scores of the others who have done the same test. In Figure 11 (page 37), this is when their total scores constitute a norm, helping in estimating their proficiency levels.

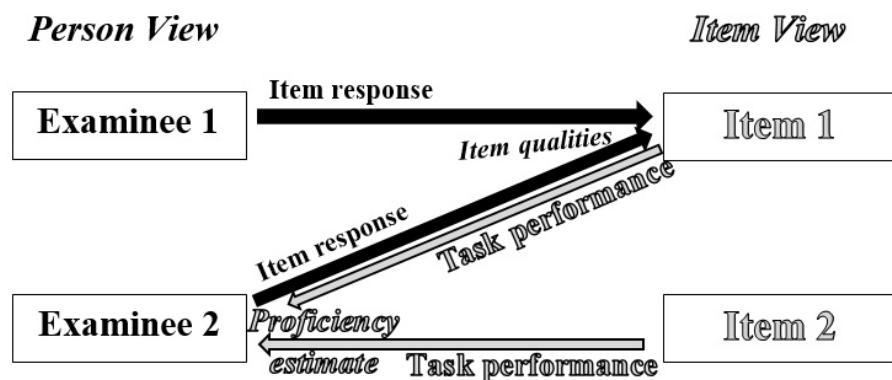


Figure 12 Example of person-item relationship in ConCloze testing

A main purpose of ConCloze is to serve as a proficiency test (page 23). In light of the person view described, the test purpose can be related to the current framework for construct measurement. Namely, the test takers' total scores, again, can be compared with one another, and so the proficiency estimation can function as test feedback for the purpose. On the one hand, this validity study does not emphasize giving feedback to the respondents but only treats it as part of the test logistics (page 23). Still, being able to use psychometric measures like ConCloze for assessing the examinees' proficiency is central to language assessment. In fact, because this study focuses on a ConCloze format and eventually on the item type, it is psychometrically sound to assume that each respondent would possess a relatively stable level of the construct competence throughout each test session. Otherwise, their collective test results could not be deemed to provide a valid norm for proficiency comparison as the proficiency reflected in one item response would instead be regarded as belonging to a different area of proficiency in another item response.

Complementary to the person view, the second way that norm-referencing is used in this validity investigation is the item view. Exemplified in Figure 12 above, item responses from ConCloze examinees can be used for comparing the items in terms of their qualities. For example, in each testing, a group of respondents could deal comparably with tasks of the same item type. Yet, easy items are likely to be tackled

successfully by most of them whereas difficult items may be scored by only those with high proficiency. In this way, the items are assigned differing difficulty levels depending on the success rates obtained from the data set of item responses.

Considering the item view, three aspects of appropriateness can be set out in favor of the norm-referenced assessment paradigm. The first is appropriateness in terms of order of construct investigation. Assuming the innovativeness of the item type, a score meaning specific to it is still unknown, let alone item properties such as difficulty and discrimination (cf. also pages 21f. for the superordinate domain taking the role of its temporary score meaning). On the contrary, setting particular criteria for score evaluation—e.g., a ‘highly proficient’ tag for over 70% of a maximum possible score—implies that these criteria should be appropriate, *inter alia*, for the target population of examinees *and* that the item properties must have been known (cf. Henning 1987: 6–7 for a cut-off score as a criterion that must be set in advance of the instruction). This means that knowing specific construct-relevant domains and justifying the score levels that constitute a satisfactorily high proficiency level and those that constitute a low one would be prerequisites to referencing evaluations of task performance to criteria – a situation that is not applicable in this foundation study. Accordingly, criterion-referencing would be unsuitable, and norm-referencing is chosen as the main assessment paradigm.

Related to order of construct investigation, a second aspect of appropriateness of the norm-referenced assessment paradigm is in terms of the purpose of application. To begin with, the focus of the current study is on test interpretation. By contrast, criterion-referencing can be deeply intertwined with test use, which in turn informs what criteria would be suitable for the purpose (cf. also Fulcher & Svalberg 2013: 5 for a theoretical description or data about task performance functioning as essential bases for criterion-referencing). For example, administering a test for assessing learning achievement may need to give out a pass-or-fail outcome through a cut-off score criterion that is justifiable. When using the same test for placement purposes, a multi-tiered system may be involved instead, as the examinees are to be categorized into different groups. Consequently, being unable to map out a specific area for test use and hence also unable to pinpoint which levels should be considered proficient in that context of application means that criterion-referencing would fall short of meaningfulness in categorizing ConCloze-score evaluations.

In light of the necessity to have a specific purpose for using criteria, it seems that norm-referencing would be relatively more suitable for the purpose of application in this study. It is worth restating that a careful approach to validity investigation has been laid out through iterative design (page 39). By looking through the lens of norms for multiple test versions administered to different groups of examinees, those norms could already help to co-construct a meaningful test-score interpretation for the item type (cf. Kane 2012b: 8 for referencing scores to norms of different groups for construct definition). For example, two large quantitative test versions that are based on different sets of items (ConCloze 5–6) will be investigated so as to seek evidence of intersecting regression lines, which indicates centrality (i.e., norm) across different facets of construct measurement (pages 206f.). In sum, given the research purpose of ConCloze construct definition, interpreting responses to each item in relation to those of the others in the same administration could appropriately fit the purpose. The norm-referenced paradigm will become even more helpful considering the research design in which the score meaning is co-constructed by multiple test versions.

Truly representative samples of the non-native English-speaker population could be huge and demand excessive research resources (cf. Section 3.2.3 for population and sampling). Moreover, the boundary of the non-native population is also difficult to define precisely (Crystal 2003: 69ff.). Accordingly, estimating a set of score criteria that would be truly reflective of and thus suitable for the population may be deemed impracticable in this research context. On this account, the last aspect of appropriateness of norm-referencing is feasibility. Evaluating the items based on the item responses that are actually elicited for research purposes offers a practical option on estimating item properties, which could be promising considering otherwise attempts to do so with a criterion-referenced paradigm for accurate population representativeness.

Apart from the appropriateness of norm-referencing for item evaluation, the assessment paradigm could also be useful for both construct definition and future studies. For example, in defining the construct, ConCloze item components may be deliberately adjusted. If item difficulties change accordingly, then the item components modified are likely to be processed for task completion. Hence, even though accrued and analyzed by means of norm-referenced approaches, the evidence of such varied item difficulties could be valuable when used collectively for describing the construct-relevant areas. In other words, when the item spec and particular changes to it bear systematic results across the

item responses—and hence the norm made for those features—such changes can be said to create variations as theoretically expected and thus provide substantive-validity evidence for construct definition (Messick 1993). Illustrated in Figure 11 (page 37), this is when the patterns of item responses observed eventually lead to making inferences about (a) the effects of changes in ConCloze-task content as well as (b) the examinees' behaviors in processing word associations (to be detailed in Chapter 3 and Chapter 4). On this account, it seems that norm-referencing is both appropriate and useful for this validity investigation in several aspects. As to usefulness of norm-referencing for future studies, it will be collectively discussed in the recommendations for future research in Part 5.3.

Another area of usefulness of the norm-referenced assessment paradigm is for future studies. In light of the utility in modifying item components, the test specs provided would be able to serve as blueprints for research replication. Equipped with tangible results of specific item designs from this study, ConCloze of alternative specs could be generated informedly (to be detailed in Part 5.3). While different samples in those research replications could probably yield slightly varying results depending on the sample characteristics, doing so with validated score interpretations as well as empirical results from this study could be easier and more resource-efficient than conducting them uninformedly.

In sum, the item responses are used in the item view as a primary tool to produce norms for empirically describing ConCloze item qualities (cf. Wilhelm 1996 for a holistic assessment model combining, among other things, norm-referenced assessment with elements of criterion-referencing). Taking into consideration the innovativeness of the item type (page 14), the current framework for construct measurement seeks to let the data speak for themselves by following a norm-referenced assessment paradigm in interpreting both the task performance and item responses. The limitations of this paradigm will be discussed in Part 5.2 (pages 275ff.).

Earlier in this part, multicomponentiality has been hypothesized to be potentially prevalent in ConCloze item prompts (pages 35f.). Given this, a specific linguistic category, e.g., lexemes, multi-word expressions, also needs to be justified for the KWIC position (cf. Figure 1, page 1 for an example of a KWIC position, in which the blanks represent an individual word serving as the question word of the task). The aim is to ensure comparability of task content, such that norms for item behaviors can be

consistently established out of item responses. For the purpose, individual words are to serve as target units because words are bases for expressing virtually everything in a language (Lewis 2000; Read 2012: 307). Irrespective of what language components that may be found processed during task engagement, selecting a right option word would be a linguistically appropriate task because words could most likely connect to other aspects of the language retained in the concordance lines. Illustrated in Figure 11 (page 37), selecting a right KWIC is hypothesized to be necessary for engaging in the test-task content. A speculation is thus that multiple language components are also likely to invoke their associations with the key word in the node position of each item prompt (to be detailed later). The item responses may thus show what exactly the components are through their patterning. For convenience, *key*, *key word*, *key KWIC*, *KWIC word*, *node word*, *right/correct option word*, *question word*, *target word*, and *target unit* will be used interchangeably, depending on the context and emphasis and unless specified otherwise, in referring to the word that is the correct answer of a ConCloze question, as in *finish* of Figure 1 (page 1).

Given that individual words are determined as target units for ConCloze tasks, a concept in assessing language ability is also worth considering. In vocabulary testing, simply because one particular word is known to the learners does not guarantee that another must always be known to them (Schmitt 2010). This could be true to the extent that two words from a close frequency level may not necessarily be similarly known. Therefore, sampling words for vocabulary testing needs to comply with certain criteria, insofar as, for example, the number of known words in a target language can be estimated reliably (e.g., vocabulary size estimated by Schmitt *et al.*'s (2001) Vocabulary Levels Test).

Considering the concept of vocabulary testing, a different portrait of language assessment can rather be seen in this framework. Having set individual words as the target units of ConCloze tasks, other elements of task content such as the genre of concordance lines will be determined later in this part. Deemed only part of the task content designed for the KWIC position, selecting individual words does not automatically mean that the item type would be presumed to be a vocabulary test in which whether each option word is known or not needs to be determined precisely. In fact, the purpose of the test at this stage of research is still to serve as a proficiency test on grammatical and vocabulary use (cf. page 23 for details). Given this proficiency testing on language use, words can thus

be considered part of the medium for estimating the proficiency levels: the more ConCloze tasks the examinees can grapple with successfully, the more likely they can be safely described as proficient in the language use. A comparative example is when a user with high proficiency in general English use will likely be able to deal with a variety of test tasks, with a higher success rate than the one with a low level of the same domain of proficiency. The limitation of this approach to proficiency testing will be discussed in Chapter 5 (Part 5.2, pages 278f.).

In light of the definition of the hypothetical construct proficiency, it is worth highlighting that even though commonality of the construct proficiency is assumed earlier (page 39), the notion of differences in the number of tasks that the proficient and less proficient users can deal with successfully is not in contradiction to it. This is based on the ground that just because an examinee has a low construct proficiency does not mean no proficiency altogether. Each ConCloze item functions as a probing tool in estimating how well or poorly an examinee may be ranked in the construct being defined. As such, when compared with the normative performance result that all of the examinees have co-constructed under the norm-referenced assessment paradigm, the performance level of that examinee can be marked with a meaningful construct-proficiency interpretation (Proficient or inept? In what language domain or set of domains?).

Considering that the test purpose has been clarified in light of the linguistic category for the KWICs, a related note on option words is also worth mentioning. Listed earlier (page 44), a correct option word may also be called a *target word*, a term that is generally used in vocabulary testing (e.g., Schmitt 1999; Read 2000). That said, again, the purpose of the ConCloze test developed is conceptualized to be a proficiency test, rather than a vocabulary test (cf. page 23 for details). Hence, the use of *target word* is for convenience in this study as the term *key* is often understood to refer to one of the options that is the correct answer to a multiple-choice test question but does not emphasize whether or not it is an individual word. On this account, (a) that the concordance prompt in each test question revolves around a target word, and (b) that the test task requires that an option word be selected for the KWIC position do not mean that the test would necessarily become a vocabulary test.

In Chapter 1, the ConCloze construct domain involves professional and academic English (page 23). Considering this, the choice of a genre—whether it is a professional

or academic one—also needs to be decided on. This is in order that, apart from individual words selected as the target units in the KWIC position, a specific genre can likewise be consistently encompassed in the concordance prompt. For the purpose, the academic genre is selected over the professional genre for two following reasons.

The first reason for choosing the academic genre is practicality. First of all, many corpora available online seem to have an academic genre or equivalent (e.g., International Corpora of English (British and Australian components), Davies' (2008–) Corpus of Contemporary American English (COCA)). By contrast, not many would have a professional English genre dedicated as such. Accordingly, the concordance prompts in this study are made of concordance lines retrieved from the Academic Subcorpus of COCA. An ever-growing text compilation for corpus-based studies of a mainstream variety of English (American English, cf. Crystal 1997, 2003 for a unique place of American English globally), COCA is one of the largest corpora available presently. Sourcing test-task content from COCA can be useful because this should allow ample diversity in text sources such as academic journals across many disciplines, thereby enhancing representativeness of the texts in real language use. On this account, sourcing concordance lines from a readily available online academic corpus may be deemed a practical option in genre selection. Details about developing the items will be provided in Section 3.2.2 and in the test spec of each subsequent ConCloze version.

A second reason in support of selecting an academic English genre over the professional one is the scope of potential inclusion. In applied linguistics, a general consensus is that English for academic purposes can be either a distinct discipline from English for professional purposes or a subset of it (summarized in Ruiz-Garrido et al. 2010: 1–2). In deciding between academic English and professional English, selecting the former genre would thus be a relatively suitable choice in this research context. The reason is that it would either reflect the purposes of the CAE program, in which following a university course is an important facet (cf. Figure 9, page 21), or also include elements of the latter genre. In whichever case, measuring the construct proficiency through sampling texts from an academic genre seems to provide a coherent ground for score interpretations of the ConCloze item type.

In formulating the research questions, it has been suggested that the literature on pedagogical use of concordances lack consistency in the substantive domains proposed

(pages 16f.). Accordingly, no meaning of existing concordance-based materials can be coherently applied to the current testing. Moreover, it has been argued that presuming legitimacy of CAE's linguistic components for this validation could be precarious because its concordance-based format is still significantly different from that of ConCloze (page 33). On account of these substantive deficiencies, it seems difficult to select a criterion test purely on the basis of the hypothetical construct (page 22; cf. also pages 18f. for the research question on the external aspect of construct validity). In fact, given the innovativeness of the ConCloze item type, it would be epistemologically paradoxical to find a validated criterion test that would measure exactly the same domain or the same set of domains of competence as ConCloze does. For these reasons, while efforts have been put into seeking a criterion test whose score meaning would be close to the hypothetical construct of ConCloze, it is determined that a criterion test for this study should aim for construct empiricism, rather than construct fidelity. This means that the purpose of administering the criterion test is to give construct-related information, which may not necessarily encompass all aspects of the hypothetical domain of competence.

In light of the construct framework outlined thus far, a criterion test can be selected meaningfully. Read's (1993; 1998) Word Associates Format (WAF, aka. Word Associates Test) is selected for four reasons. First, a validated version of the test is readily available for downloading, which is at <http://www.lexutor.ca/tests/associates/>. One item is illustrated in Figure 13 below, where the word *favorable* functions as the target word (equivalent to a KWIC word in the ConCloze task). Four words in the two columns underneath are associated with *favorable* and are required to be selected for task completion (answers: *helpful*, *positive*, *response*, and *weather*) (to be detailed later). The second reason is its simplicity in form (Read 2012). This potentially makes an administration of the test practical and undemanding of the research resources.

Question 19

favorable

- | | |
|-----------------------------------|-----------------------------------|
| <input type="checkbox"/> helpful | <input type="checkbox"/> habit |
| <input type="checkbox"/> legal | <input type="checkbox"/> response |
| <input type="checkbox"/> possible | <input type="checkbox"/> teacher |
| <input type="checkbox"/> positive | <input type="checkbox"/> weather |

Figure 13 A WAF item (Read 1998, cited in Cobb ca. 2011, adjusted)

The third reason for selecting Read's (1993; 1998) WAF lies in its previous validation studies. Read (1998) piloted a WAF version on 16 undergraduate Japanese students and, *inter alia*, administered two alternative forms of the test on approximately 100 international students at the undergraduate level. Read (2007) mentioned five works about different versions of the test, mostly related to undergraduate Dutch students. Schmitt (2010: 226ff.) also mentioned several identical works related to different versions of the test that were carried out in Northern Europe. Schmitt *et al.* (2011) investigated how to interpret scores from different versions of the test on two groups of 44 international university students in total, the majority of whom had Japanese and Chinese as their L1s. Lastly, Batty (2012) performed a factor analysis of item responses from 530 Japanese university students. In light of this review of some of the previous studies into WAF, it seems that, on the whole, the respondents used in relation to WAF could be argued to have no exclusive tie-in to one particular L1 background.

Judged from the number of related studies, previous research into WAF may be deemed fairly extensive. On the one hand, as Schmitt (2010: 228) cautioned, those studies use different versions of WAF. Also, the score interpretations are still uncertain in terms of the chance of variance from guessing in the test tasks (*ibid.*; Schmitt 2011). The many forms of task content and the chance of guessing may thus be considered drawbacks to test-score interpretations. This is so because a lack of uniformity in task content would result in a lowered power of extrapolation to the construct competence that those studies could be claimed to commonly test (cf. page 23 for a similar idea about a trade-off between extrapolability and generalizability of test-score inferences). On the other hand,

however, their score interpretations are based on a substantial number of observations admissible in the performance universe. Equipped with a collective power of generalization, this means that the central idea of WAF measuring word associates (to be detailed later) could be reasonably representative of the capability of WAF items.

The discussion thus far seems to suggest that the WAF-score interpretations are based on extensive previous validations of the score meaning. Accordingly, higher confidence can be given to the generalizability of their construct inferences than to that of a criterion test which has few validation studies to support. Given that none of the studies in the earlier paragraph found a core domain tested different from knowledge of word association, it seems safe to say that, notwithstanding limited extrapolability, the primary WAF score meaning is very likely to lie in the domain, which will be clarified later. In fact, it is also worth highlighting that in spite of many different studies into WAF validations—from Read (1993) to Schmitt *et al.* (2011), from Read (1998) to Batty (2012)—it is the central score meaning, rather than each test version itself, that matters for an evaluation of the score-meaning generalizability. The same is true with the rationale for investigating the construct validity of the ConCloze item type, which has been discussed in Chapter 1 (pages 2f.).

The last and most important reason for choosing WAF as a criterion test lies in its construct competence. To begin with, it has been argued earlier that an individual word, not a formulaic expression, for instance, seems to be suitable for functioning as the target unit of each ConCloze task (page 44). In a WAF task, an individual word is also the target word, as in *favorable* of Figure 13 above, which may be equivalent to a target unit in a ConCloze task. On this account, both the ConCloze and WAF formats seem to have individual words as the target units with which other information would need to be associated. Further, ConCloze respondents are given multiple concordance lines that one of the choices could go well with. WAF respondents are likewise required to process different words with which the target word has association. For example, in Figure 13, four words on the left column and another four on the right are different pieces of information that the respondents have to process in order to select four among them as the most appropriate and well associated with the target word *favorable*. This means that both of the formats seem to require their respective respondents to process multiple pieces of information and select the word(s) that would bring about the greatest compatibility with them. In sum, both the ConCloze and WAF formats are argued to call for a

comparable substantive aspect of task engagement: processing different pieces of information and matching them to a given word.

In addition to the substantive aspect informed by the task requirements, the literature on WAF validity also suggests several aspects of substantive content potentially suitable for ConCloze construct empiricism. First of all, WAF scores are found to vary significantly in accordance with the students' proficiency levels (Read 1993). The purpose of the ConCloze test is also to assess proficiency in language use (cf. page 22 for details). This indicates that WAF could potentially be used as a scale the performance of which can be hypothesized to vary positively with ConCloze scores. As such, for example, a high performer in WAF can be expected to likewise have a high proficiency level of the ConCloze hypothetical construct.

The second reason of substantive content for ConCloze empiricism lies in the semantic relations formed by words in task content. Each question word in WAF can be considered the node of a lexical network with the associates related to it (Read 1998: 56; Schmitt *et al.* 2011). For example, *favorable* of Figure 13 above may be deemed the node of the network in this item, and *helpful*, *positive*, *response*, and *weather* are its associates. Similarly, the ConCloze format could be regarded as based on an individual word functioning as the node in the KWIC position (pages 4f.). It thus seems reasonable to presume that the aspect of vocabulary use subsumed in the ConCloze construct is when the lexical network in the concordance prompt would give a hint about a right KWIC for task completion (cf. a superordinate domain as the full hypothetical construct on pages 21f.). Accordingly, incorporating WAF as the criterion test could potentially validate if there is such a network that operates in ConCloze. Also, it is worth differentiating two terms discussed frequently: *associate* and *association*. The term *associate* will be used mainly in referring to an individual word that is related to another linguistic unit, usually another word, as in *word associates* in WAF. The term *association* will be used as a generic term, referring to individual words as well as other linguistic units that can be deemed frequently co-occurring with another. For example, elsewhere in this thesis, collocations, colligations, and semantic prosodies could all be considered *associations* of KWICs.

Apart from the previous works about varying according to language proficiency levels and about a question word forming a lexical network, a third point from the

literature supporting WAF as the criterion test is about word meaning in task engagement. Core elements of language measured in WAF are empirically found to be those of the meaning of a target word (such as something supportive in *favorable* of Figure 13 above; cf. Read 2007: 113; Schoonen & Verhallen 2008 for aspects of meaning of a target word processed). Likewise, the ConCloze format offers word options to be chosen for task completion, and their core components in meaning could be those which the examinees would process and decide upon. In fact, categorized in Table 3 (page 34), word use in authentic language would usually engage other frequently co-occurring words such as their collocates. This means that the lexical network discussed previously may likely consist at least of collocates of the right KWIC and give a hint about it potentially by means of its core semantic components. Therefore, pairing WAF to the ConCloze validation could shed light on whether the semantic components of the ConCloze options will be important for task completion. This would then allow for specifying the construct-relevant domains as intended in formulating the hypothetical construct (Section 1.3.2).

Latest works supporting WAF suitability as the criterion test involve the semantic relationship among the words in the item. Batty (2012) found that general vocabulary knowledge accounts for most of the WAF sub-item variances whereas knowledge of synonymy and knowledge of collocation additionally explain separate sub-items on their respective dimensions. Further, Read (2012) described that the lexical network which is centered on a target word may be formed by associates that are in either a paradigmatic, syntagmatic, or analytic relationship with it, e.g., synonyms, collocates, whole–part, respectively. A current implication is that given the inherent multicomponentiality of ConCloze (pages 33f.), it is likely that WAF and ConCloze item responses will be at least correlated with each other with moderate strength. For this study, this could be particularly promising considering the multiple language components hypothesized earlier to be retained in the concordance prompt (page 38). Accordingly, having WAF as the criterion test would mean randomizing those types of semantic relationship in item responses and hence, for example, a strong correlation between the two tests could imply a leading role of semantic associations at work—particularly as formed by knowledge of synonymy and knowledge of collocation—in ConCloze task engagement.

In conclusion, Read's (1993; 1998) WAF is selected as a criterion test for four reasons: availability, simplicity in form, extensive previous validation studies, and its substantive comparability. Regarding the last reason about substantive content, the

measure is deemed potentially suitable for ConCloze construct empiricism because the examinees of both the WAF and ConCloze formats would be required to process different pieces of information and match them to a question word. Moreover, the literature review indicates that both of them are likely to have (a) a positive relationship with language proficiency levels, (b) a lexical network formed by their respective question word, and (c) semantic elements of their question word(s) processed and decided upon for task completion. Because WAF randomizes three types of semantic relationship in item options (discussed earlier), it is hypothesized that at least a moderate association between the two tests could be observed, given the assumption that multiple language components—collocates and synonyms included—are likewise encompassed in the ConCloze concordance prompt.

Given that a criterion test has been selected for ConCloze, a final note would be on a contrastive example of another measure that may otherwise serve as a criterion test. Illustrated in Figure 14 below, Gyllstad's (2007; 2009) COLLEX (COLlocating LEXis) and COLLMATCH (COLLocate MATCHing) tests ask the examinees whether strings of two–four words are real collocations in the English language. The test formats are not selected as a criterion test for two reasons. First, because those formats are relatively new when compared with Read's (1993; 1998) WAF, their development process still has a tie-in to respondents of only one L1 background in validity investigation (i.e., Swedish). This means that even though the findings indicate a content-validity aspect of collocational knowledge—a competence domain of language use potentially construct-relevant to ConCloze (cf. pages 33f.)—their construct interpretations have lower representativeness of the capability of their corresponding items (i.e., limited generalizability) than WAF. Secondly, as their names suggest, COLLEX and COLLMATCH tap into knowledge of collocation exclusively. By contrast, WAF scores tap into at least knowledge of synonymy and knowledge of collocation. Accordingly, as the concordance prompt is hypothesized to involve multiple language components (pages 35f.), scores from COLLEX and COLLMATCH are less likely to co-vary with ConCloze scores than WAF scores are.

COLLEX

			A	B	C
1. do damage	make damage	run damage	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. turn out a fire	put out a fire	set out a fire	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

COLLMATCH

- | | |
|------------------------------|------------------------------|
| 1. have a say | 2. lose sleep |
| <input type="checkbox"/> yes | <input type="checkbox"/> yes |
| <input type="checkbox"/> no | <input type="checkbox"/> no |

Figure 14 Gyllstad's (2007; 2009) COLLEX and COLLMATCH

In conclusion, the concordance-based task content, an essential component of each ConCloze prompt, is argued to accord with the original framework of the CAE program and have multiple language components and discursive interrelatedness. The two qualities are presumed to be retained in the testing situations, making the item type a potentially efficient and construct-relevant measure of professional and academic English grammatical and vocabulary use. Based on the item feature of multiple corpus-retrieved lines, some interrelated aspects of usefulness are hypothesized as belonging to the measure: reflecting real language use, sampling texts comparatively thoroughly, neutralizing authors' idiosyncratic features, lessening domino effects on inter-item dependence, and being more usefully representative of language use.

Considering the innovativeness of the task format, another conclusion is to choose norm-referencing as the primary assessment paradigm, which is coupled with its appropriateness in terms of order of construct investigation, purpose of application, and practicality for construct definition. In letting the data create norms for themselves, individual words are selected as question units in the tasks, serving as part of the medium for estimating proficiency levels of the language use. Further, for practicality, Davies' (2008–) COCA subcorpus of the Academic Genre is selected for sourcing test-task content. For construct empiricism, Read's (1993; 1998) WAF is selected as a criterion test because of (a) its availability, (b) simplicity in form, (c) previous validation studies without any exclusive focus on respondents of one particular language background, (d) similarity in substantive content informed by its task requirements, and (e) previous

findings on its positive relationship with language proficiency levels and on semantic components of a question word forming a lexical network with multiple types of associates.

2.4 Lexical Priming in Cloze Procedure

In ConCloze, the KWIC position of an item prompt features a column of gaps, which represent the word asked in each question (as in Figure 1, page 1). This could be deemed a feature derived from cloze testing—the words being questioned are *clozed*. Filling out the gaps in a cloze test can be equated with visual closure, in which the human mind seeks to find a meaningful pattern contained in a whole frame (Ohnmacht *et al.* 1970; Oller & Conrad 1971; Zinkhan & Martin 1983). For illustration, an example in Figure 15 below is worth considering, in which every five words are deleted (answers in parentheses) (cf. Brown 1980; Alderson 1980; Lange & Clausen 1981; Chapelle & Abraham 1990; Read & Chapelle 2001 for differing terms for cloze types).

Words are missing in the blanks below. Fill them in with one appropriate word each.

An issue that arises in contrasting ConCloze with Butler's test is a lack of contexts. A counterargument goes that (ConCloze), whose prompt is made (of) a concordance, is void (of) contexts. For example, the (prompt) is reprehended for being (so) fragmented that macrostructural linguistic (features) such as author's style (and) intention, discorsal cohesive devices (would) become entirely incomprehensible to (the) examinees. This counterargument is a side issue, though, as use of texts full of contextual information does not automatically guarantee nor disqualify the validity of an item type.

Figure 15 Example of a cloze test

Doing cloze tests is claimed to involve pragmatic expectancy grammar (Purpura 1999). This refers to “any [*integrative*] procedure or task that causes the learner[s] to process sequences of [language] elements... that *conform to the normal contextual constraints*... and which requires the learner[s] to relate sequences of linguistic elements via pragmatic mappings to extralinguistic context” (Oller 1979: 38, italics mine). In other

words, the cloze procedure operates when a gap is presented in the surrounding co-text. An examinee will try to fill out the gap using the restrictions imposed by its co-text, oftentimes by means of what can normally be expected extralinguistically. Also, it is worth considering that the current discussion separates the contexts of a gap into two categories: the co-text surrounding a gap, and the contexts beyond the co-text. Separating the contexts this way can be useful because it accords with the ConCloze situation where the co-texts are exclusively those contained in the item prompt (cf. also a related note on *co-text* vs. *context*, page 4).

In light of the cloze procedure as defined by Oller (1979), a few key concepts deserve elaboration and will be applied to ConCloze in turn. A first concept is that the cloze procedure is an integration of processes and of multiple domains of competence. In psychological terms, this is also called crystallized intelligence—a neutral “complex of abilities” formed by experience, education, and acculturation (Ackerman *et al.* 2000: 108). Secondly, the examinees would predict what the writer could have meant to say from the text that remains; they would then seek to fill out the gap in the way that still maintains the co-textual normality (Oller 1979). In other words, the examinees would avoid breaking the normality governing the text that contains each gap. Lastly, the cloze examinees would also draw on a map of such normality that takes into account the conventions beyond the co-text. These conventions would include, but not be limited to, cohesive devices, genre features, author’s intention, and discourse structure.

In light of the concepts discussed in relation to cloze testing, some implications related to the ConCloze construct can be as follows. First, the ConCloze item type has been hypothesized to operate using multiple language components in the concordance prompt and have the hypothetical construct of proficiency in grammatical and vocabulary use (pages 35f. and 22, respectively; cf. also Figure 1, page 1 for item components). Thus, assuming the cloze procedure operates in ConCloze, the gaps in each prompt are likely to call for the examinees’ integrated ability to use grammar and vocabulary. Depicted in Figure 16 below, this is when multiple aspects of language in the prompt would be drawn on as a concerted activity. In fact, the cloze element is empirically found to involve assessing language proficiency (Hanania & Shikhani 1986). Accordingly, not only does the cloze procedure match the multicomponentiality argued to be characteristic of the item prompt in ConCloze, but the procedure also supports the idea that the test purpose

could be reasonably designated as a proficiency test (cf. page 23 for setting the test purpose).

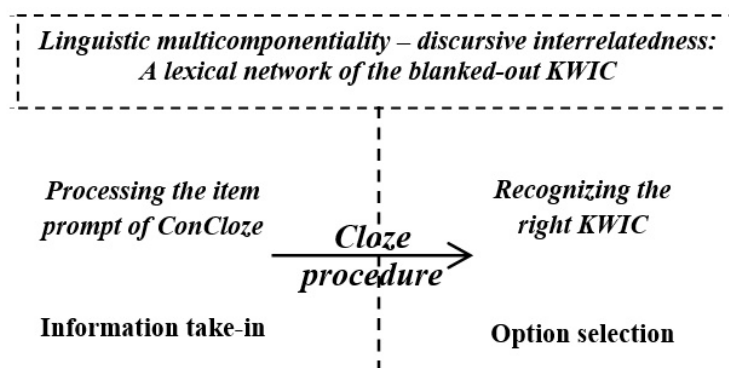


Figure 16 Cloze procedure in ConCloze processing

A second implication for ConCloze carried by cloze-related concepts is about textual normality. In ConCloze, linguistic features such as collocation and constraint on language use are hypothesized to be present in the item prompt (pages 33f.). Then it is also conceptualized that a lexical network could be formed in the concordance prompt, centering on the key KWIC (page 50). In light of the cloze procedure, in which the examinees would seek to fill out the gap in the way that still maintains the co-textual normality (Oller 1979), it may be proposed that those linguistic features could be part of the normality that ConCloze examinees process and use for task completion. In other words, the construct competence is likely to be stimulated by the lexical network that could be normally expected in the vicinity of a blanked-out KWIC.

In light of the textual normality in the form of frequently co-occurring linguistic features, an interpersonal dimension of the ConCloze construct is also worth considering. Introduced earlier, the cloze procedure has a partial basis in experience and education, which will definitely vary from person to person. Accordingly, another point in relation to the textual normality is that the ability to use grammar and vocabulary may likewise vary accordingly in ConCloze. This implies that the examinees' engagement in ConCloze may be driven by combining different aspects of grammatical and vocabulary use that would suit them best individually. The interpersonal variations may thus result in

differing total scores reflecting varied proficiency levels as hypothesized earlier (page 39).

The last implication for ConCloze carried by cloze-related concepts is about contexts of application of the cloze procedure. In cloze testing, one word is usually required for each blank (as in Figure 15, page 54). However, in ConCloze, one word is required for multiple blanks in the node position of the same item. Should ConCloze be found operational when actual item responses are analyzed, then a new context of the cloze application will be suggested. This would then allow an inference that not only can the cloze procedure take place in a passage with multiple blanks, one word for each, but the mechanism could also occur in another kind of text called ConCloze, in which multiple blanks may confer synergy by forming a lexical network for task completion. In sum, reviewing the cloze procedure in pragmatic expectancy grammar allows the implications that (a) the proficiency in grammatical and vocabulary use in ConCloze are likely to be integrative in nature, (b) the lexical network hypothesized to be formed during ConCloze engagement is normally expected in the context of the KWIC, and (c) the operationality of ConCloze would be a new context of application for the cloze procedure.

In arguing for the pragmatic expectancy grammar, Oller (1979: Chapter 3) alluded to internalizing co-textual and contextual norms. The internalization of the norms may be equated to learning and acquisition of the integrated knowledge used for gap filling. Therefore, the examinees' ability to perform the cloze procedure as well as their norms for language use are set by what they have learned. Nonetheless, *how* the norms become the norms for the examinees the way they are seems unemphasized in the grammar. As Oller (1979: 24, italics mine) put it, “[r]ather, we are concerned with the psychological realities of linguistic knowledge as it is internalized *in whatever ways* by real human beings.” In fact, for example, collocation—a language phenomenon which has been reviewed in this study as a type of associate that helps to give a hint about a KWIC (pages 22, 33f., 51)—seemed to be taken for granted as a ‘factive’ aspect of norms in language use (op. cit.: 17ff., 26). He viewed that the factive norms are among those that are already distinctively encoded and which the learners need to internalize. Most importantly, Oller did not explain adequately how the norms manifest as such. On this account, an argument can be that the pragmatic expectancy grammar assumes what the learners have learned as norms for filling out the gaps, but does not stress how they can distinguish between idiosyncratic co-texts and contexts and the norms in testing situations. In terms of

internalization, this means that norms are not clearly differentiated from styles of individual authors and the specific content which vary from text to text.

In light of the inadequacy of the pragmatic expectancy grammar, two theoretical issues arise against ConCloze engagement. The first issue is about recognizing the norms: this study presumes that the examinees have never seen the test items before they take the test. If all the linguistic features in the item prompt were completely unrecognizable to the examinees, then they would not be able to complete the test task successfully. Despite the unseenness of the prompt, there must be linguistic normality that serves as a link to the blanked-out KWIC. Presuming that the hypothetical construct (pages 22f.) is reasonably accurate, the link would be the language components that habitually co-occur with the KWIC and can give clues about it. In other words, the pragmatic expectancy grammar could only be used for assuming what has been learned and internalized would serve as a norm for the expectancy in ConCloze. The grammar cannot explain adequately how the ConCloze examinees could recognize the clues in the snippet texts of the item prompt. Without recognizing the norms as they appear in the concordance prompt, the examinees could not differentiate the target words from their distractors, either.

The second issue against ConCloze engagement is about the habituality of the multiple language components co-occurring with the KWIC. The pragmatic expectancy grammar could not explain adequately how a linguistic feature could become commonly expected as habitually co-occurring with one particular word. It is unlikely that all the words in the prompt would be equal in terms of ‘hintingness’ towards the blanked-out KWIC. For example, a group of examinees would recognize that a set of clue words, when seen in the same context, will more likely be associated with one word rather than the others. An example is when the words *recent*, *extensive*, *conduct*, *systematic*, *investigation*, *journal*, and *findings* seems to invoke into cognitive processing the word *research*, rather than *newspaper*. Also, the inadequacy of the grammar is especially true considering that clue recognition must be shared by multiple examinees for a validity claim – an assumption for the operationability of the construct competence (page 39). Without a shared pattern among them, the item responses would then be invalid for inferences about a competence. In sum, the pragmatic expectancy grammar cannot adequately account for how the ConCloze examinees have internalized the habitually co-occurring words of a particular word and would be able to separate them from idiosyncratic features in the concordance context. Most importantly, the grammar cannot

explain adequately how the ConCloze examinees could end up sharing these patterns necessary for task engagement.

To explain the genesis of ConCloze processing, this review adopts Hoey's (2005; 2013) theory of lexical priming as a theoretical framework underlying task engagement. Originating from corpus-based insights into lexicogrammatical patterns, the theory is deemed "a sound explanation of why collocations, colligations and semantic associations exist" (Pace-Sigge 2013: 168). In priming words and their associations, learners learn not only individual words in every encounter, but also learn a variety of associations in relation to the words. This latter kind of learning is priming, in which "what is primed to occur is seen as shedding light upon the priming item rather the other way round... every word is mentally primed for collocational use... loaded with the contexts and co-texts in which it is encountered" (Hoey 2005: 8). For example, *validity*, *investigation*, *concordance*, *cloze*, *research* and *language testing* could have been primed at this point for the coinage *ConCloze*. *Investigating construct validity* and *in a dissertation* would be primed for the overall context. Priming is subject to (a) strengthening by means of repetitions of similar encounters and (b) weakening by means of unfamiliar contexts and/or individual overriding. In sum, "everything we know about a word is a product of our encounters with it" (ibid.: i). Lexical priming is incorporated into ConCloze processing in Figure 17.

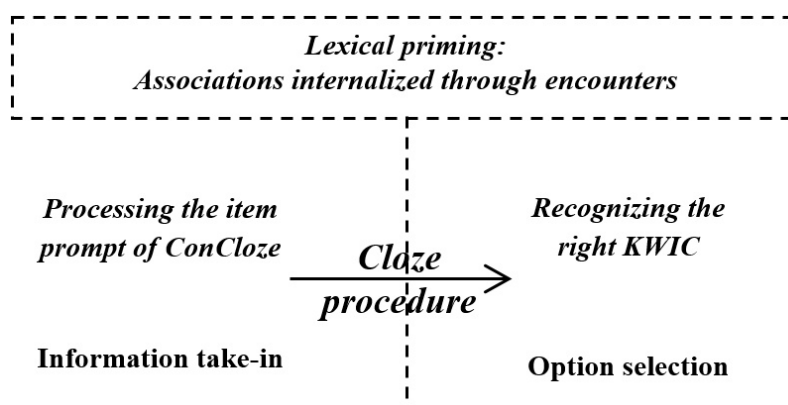


Figure 17 Lexical priming in ConCloze processing

In light of the lexical priming, the theoretical issues against ConCloze engagement could be addressed as follows. First, the pragmatic expectancy grammar cannot account

adequately for norm recognition, but the priming theory views co-occurrences as natural phenomena in language acquisition and use. Certain features of a word tend to be used more associatively in the same contexts as the word than the others. Such features include words (i.e., collocates), structures (i.e., colligations), and meaning (i.e., semantic prosodies). The associations become patterns, and users acquire these patterns through repeated encounters. They in turn use these patterns in their language production, thereby priming and/or reinforcing the patterns on the other users. Accordingly, it follows that the concordance lines of ConCloze are products of the pattern use. The examinees, especially the proficient ones, would have these patterns internalized during their language learning and acquisition and be able to recognize the patterns.

The second issue that the theory of lexical priming can help to address is the habituality of the linguistic features co-occurring with a KWIC. The pragmatic expectancy grammar cannot account adequately for why those features would be recurrent and the knowledge about them could be common among ConCloze examinees. According to the priming theory, this could be explained by repeated encounters and productions. As the patterning goes on and are repeated in communicative situations, primings become strengthened and established over time. When the learners encounter words, they also take in co-textual and contextual information. To this extent, the learners of the same language acquire not just words but also these primings subliminally yet comparably across individuals; their mental lexicons are tagged with similar contextual information.

As a final note, each KWIC tested and its co-textual associations have been hypothesized to be interrelated (pages 51f.). This hypothesis seems to also accord with the current theoretical discussion of lexical priming. Crossley *et al.* (2011) reviewed that the mental lexicon of advanced learners is generally more densely woven than that of non-advanced learners. Knowing words and their associations can also discriminate learners of different proficiency levels (Hargreaves 2000; Taylor & Barker 2008). Assuming the proficiency in grammatical and vocabulary use is reasonably accurate for the ConCloze construct, an inference would be that primings in the high performers' mental lexicon would be richer than those of the low performers. Likewise, high performers would have more of such associations between words and co-texts than low performers. They thus would know and be able to recognize more of the KWIC-related information in the test items than the low performers, resulting in higher total scores.

To conclude, individual KWIC words have been argued as a medium in ConCloze for proficiency estimation (page 44). Accordingly, in light of the theory of lexical priming, it can be conceptualized that those with high proficiency would know more of the associations in each ConCloze item *and* associations of more KWIC words across the test than the examinees with low proficiency. Also, it is worth pointing out that the theory of lexical priming was formulated focusing on language learning and acquisition (Hoey 2005). If the ConCloze scores are found varying in accordance with the criterion test WAF and thus language proficiency levels (cf. page 51 for the postulate), then the operationality of ConCloze would imply that language testing is a new application for the theory.

2.5 Concluding Remarks

In this chapter, use of concordances in language testing and assessment has been reviewed. Since Butler's (1991) prototyping of a concordance-based test, there seems to have been no recognition of the potential of concordances in forming test items. The lack is especially evident considering the functions of concordances in language assessment on a broader scale, as in all the issues of Cambridge-ESOL research notes. A likelihood is that a gap exists concerning the role of concordances as part of test-task content: to date, the concordance in its typical presentation has never been systematically used as such before. On this account, the present review helps to ascertain the legitimacy of the originality claim in Chapter 1 (Part 1.2), and that this study is making an original addition to the literature on concordance use for language testing.

In the absence of previous direct studies into concordance-based testing, its substantive processing is yet unknown. Accordingly, a framework for construct measurement has also been set in this chapter for a principled validation. This starts from (a) the multicomponential nature and discursive interconnectedness of the ConCloze item prompt, and (b) the observation that its format differs substantially from the concordance-based format in a previous edition of the CAE program (CELA 2010a). Five interrelated aspects of usefulness of those qualities have been proposed as belonging to the item type: reflecting real language use, sampling texts comparatively thoroughly, neutralizing authors' idiosyncratic features, being relatively representative of authentic language use, and fitting the target hypothetical construct (page 22) as well as the original framework

of an industrial testing program like CELA's CAE. Some of the areas of improvement that the item type could make are summarized in Table 4 from both Chapter 1 and this chapter in no particular order.

Table 4 Areas of potential improvement

Cause	Improvement
Existing tests with concordances not in their accurate form	Truly concordance-based test-task content, with concordance advantages potentially offered
Constant need for item-design innovation	Score meaning validated for improving test programs on the industrial scale
Lack of existing related validation studies	Domain-specific score meaning, probably generalizable to the CAE program and other newly developed ConCloze tests
Domino effects in cloze testing	Minimized inter-item dependence in item design
Authors' idiosyncratic features in test tasks	Idiosyncratic features neutralized by the format with multiple concordance lines
Potentially biased sampling of subject matter in traditional language tests	Relatively more thorough sampling of texts through multiple concordance lines

In light of the qualities that ConCloze items are expected to have in common, an assumption is that the language aspects specific to the item type exist and could be identified with empirical item responses. Given this, a norm-referenced assessment paradigm is selected over a criterion-referenced one due to its appropriateness in terms of order of construct investigation, purpose of application, and practicality. The paradigm is also useful for both future studies and construct definition, in which item responses are used as a primary tool to produce norms for empirically describing item qualities.

In Chapter 1, the hypothetical construct, which is subject to specification and refinement upon analyzing actual item responses, is proficiency in professional and academic English grammatical and vocabulary use (page 22). In letting the item responses create norms for themselves, individual words are selected as question units in the tasks. They are to serve as part of the medium for estimating the proficiency levels of the hypothetical construct. Further, for practicality, Davies' (2008–) COCA subcorpus of the Academic Genre is selected for sourcing test-task content. Rather than for construct fidelity to ConCloze, Read's (1993; 1998) WAF is selected as a criterion test for construct empiricism because of (a) its availability, (b) simplicity in form, (c) previous validation studies without any exclusive focus on respondents of one particular language

background, (d) similarity in substantive content informed by its task requirements, and (e) previous findings on its positive relationship with language proficiency levels and on semantic components of a question word forming a lexical network with multiple types of associates.

In this chapter, the cloze procedure in pragmatic expectancy grammar is also reviewed. This allows the implications that (a) the ConCloze construct proficiency is likely to be integrative in nature, (b) the lexical network hypothesized to be formed during task engagement is normally expected in the context of the KWIC, and (c) the operationality of ConCloze would be a new context of application for the cloze procedure. However, the grammar cannot adequately account for how the ConCloze examinees have internalized the habitually co-occurring words of a particular word and would be able to separate them from idiosyncratic features in the concordance-based context. Most importantly, it cannot explain adequately how the examinees could end up sharing these patterns necessary for task engagement.

In light of the inadequacy of the pragmatic expectancy grammar, the theory of lexical priming has been brought in. Proposed in this study as a genesis of the cloze procedure, the theory is involved with words priming their associations in acquisition, storage, and production through repeated encounters. For ConCloze, certain linguistic features around a target word are viewed as tending to be used more associatively in the same contexts as the target word. Potentially resulting in patterns of language use embedded in the concordance prompt, the features could include words (i.e., collocates), structures (i.e., colligations), and meaning (i.e., semantic prosodies). Proficient ConCloze examinees are hypothesized to have more of these patterns internalized and accordingly be able to score higher than those with a low proficiency level.

As a final note, it is worth recapping the substantive mechanism of the ConCloze item type, which has been postulated in this chapter on the basis of the literature review. It is theorized that language learners have had repeated encounters of words and their normal co-texts and contexts of occurrences in real communicative situations. This results in lexical priming: the associations are learned and internalized subliminally yet comparably across individuals, with the primed associations providing information about the priming words. In ConCloze, this is when some of the associations of a target word are present in the concordance prompt and form a lexical network, which gives the

examinees a hint about it. The cloze procedure is hypothesized to occur during task engagement, in which the examinees glean the information about the KWIC and seek to fill out the blanks in the KWIC position with one of the options provided. Integrated and multicomponential, the information includes, but is not limited to, knowledge of grammatical structures related to the KWIC and knowledge of its collocates. Because proficient language users are likely to (a) have had higher exposure to the target language and thus (b) know more of the associations in each ConCloze item as well as associations related to more KWIC words across the test than the examinees with low proficiency, the overall task performance in ConCloze testing could potentially indicate their proficiency levels of the construct competence. For convenience, unless required otherwise, *association*, *word information*, and *lexical network* will be used in this study interchangeably, denoting the information imparted by any linguistic components embedded in the concordance that can be associated with a KWIC (cf. also a related note on *word information* on page 8). In a ConCloze task, such information is theorized to emanate mainly from a concordance-based item prompt, giving a hint about the missing KWIC. Accordingly, it is worth stating that as *word information* refers to concordance-based information related to a target KWIC, *word knowledge* will likewise be used in referring to an examinee's knowledge of such information that can be related to a KWIC. That is, *word knowledge* is not meant in this study to refer merely to knowing the lexical meaning of a particular word.

Chapter 3 Developing ConCloze

3.1 Introduction

The context for this research project has been set in Chapter 1. Then in Chapter 2, a gap in the literature is identified and theoretical frameworks provided for analyzing ConCloze item responses, and for explaining the substantive processing as well as the acquisition of the construct proficiency. As parts of the validity argument, the previous chapters are outlined in Table 5 as grounds for the validity investigation and theoretical backing respectively.

Table 5 From ConCloze's background to development

Chapter and Thesis Stage		Component in Validity Argument	
1. Introduction		Research grounds	
2. Literature Review		Theoretical backing	
		Data, warrants, and claims	
3. Test Development	Test Version (Web-based)		Test Length
	ConCloze 1 Quantitative prototyping		39 items long
	ConCloze 2–4 Qualitative prototyping, usability testing		5 items long
	ConCloze 5 Field-testing		30 items long

Given the innovativeness of the item type, this study also needs to develop a ConCloze test, the process of which is shown in Table 5 above as ConCloze 1–5. Representing the prelaunch phase, the development process can be deemed iterative, in which analyses of item responses in one version inform decisions made in the next versions (cf. Fulcher & Davidson 2007: Unit A6.5 for iterations in test development). For example, item responses in ConCloze 1 are analyzed, and the findings are used for conceptualizing the specification (spec) of ConCloze 2 as well as subsequent versions. The iterative design has two advantages. First, the test spec is geared towards construct representation that is evidence-centered (cf. Messick 1994; Mislevy *et al.* 1999; Mislevy & Riconscente 2005). Another advantage is that construct-irrelevant variance can also be addressed along the development process. The test is steadily improved for validity investigation as a result.

In light of the iterative research design, this chapter is also structured so as to log the development process chronologically. Each ConCloze version is composed of the test spec, response collection and analyses, and decision making as well as construct evaluation. Highlighted in Table 5 above, this means that empirical findings and score interpretations are also offered in the current chapter in support of the entire validity argument. The benefits of structuring the thesis this way are threefold. First, it demonstrates how the ideas and insights about the ConCloze construct gradually unfold from unknownness. It reflects the developmental nature of this study, showing that the construct definition is predominantly *a posteriori*. Secondly, the structuring also culminates in inductive argumentation: patterns are sought from the observed test and item responses, leading to formulating interim hypotheses and their subsequent testing (cf. Kane 1990; Kane 1992; Kane 2006 for the notion of chains of forming and testing hypotheses). Accordingly, the construct is not defined by, for example, a fixed hypothesis which would then be merely accepted by a non-significant statistic. The last benefit is psychometric. The iterative structure makes the definition increasingly sharp; the underlying domain of competence, if any, is unlikely to be set too narrowly in the first place (cf. Messick 1993 for necessity to ensure a construct does not fail to include essential parts or dimensions of competence). Over the iterations, confidence in drawing inferences about the ConCloze construct is accordingly increased.

The structure of this chapter is as follows. The quantitative prototyping is first discussed in Part 3.2. As shown in Table 5 above, the prototype is called ConCloze 1, the designation representing both the test and the investigation stage. Then in Part 3.3 for ConCloze 2–4, the prototypes of item variants (IVs) are generated based on the spec of ConCloze 1 items. These IVs are used for qualitatively investigating the competence domain and processes. Afterwards, a test with the IV features prototyped will be produced in Part 3.4 (ConCloze 5), where the items are based on the precursor items from ConCloze 1. This chapter concludes in Part 3.5 with a summary of the evidence collected and the construct inferences, the appropriateness and adequacy of which will also be appraised. At the end of each test version, the hypothetical test purpose (page 22) will also be evaluated in light of the empirical findings, thereby adjusting and refining the understanding of the construct competence.

3.2 ConCloze 1: Quantitative Prototyping

3.2.1 Rationale

In an enterprise of test development, there can be several kinds of information obtained from prototyping a new item type. This is usually in the interests of larger-scale feasibility and financial resources, to the extent that the test will not fail in actual use (Fulcher & Davidson 2007: 76ff.). Among the information types obtainable is information about viability of a test for measuring a new construct (Nissan & Schedl 2012). A key mission is to evaluate if the test-task engagement indicates a competence domain in action or merely reflects random responses to the test questions.

When prototyping a new item type, use of both quantitative and qualitative approaches is generally recommended (Kenyon & MacGregor 2012). Considering the key mission, quantitative responses are to be examined in this version for two reasons. First, at this stage of research, whether ConCloze can actually measure a domain of competence is yet unknown. Investigating quantitative responses is a pragmatic and time-efficient alternative to making construct inferences, vis-à-vis a qualitative investigation. Secondly, aggregate patterning in item responses can be sought systematically; the likelihood of testing a distinct competence can be evaluated on a scale level. This gives higher confidence than seeking construct-related patterns from qualitative item responses in isolation.

Depending on the objective, prototyping a language test can be evaluated using measures such as content coverage, reliability, and cost efficiency (Hoshino 2009: Section 1.1). Given that quantitative responses will be targeted, test reliability is to be sought in this part as it can show a construct in action on the scale level (cf. also Gabrenya 2003 for intertwinement of validity with reliability). Also called response consistency (Loevinger 1957), test reliability is when examinees' performance is reasonably consistent throughout the test. For example, an examinee who is fully proficient in English listening would perform well consistently from beginning to end of a listening test. For ConCloze, consistent variation in item responses across individuals may thus indicate varying levels of the intended proficiency they have acquired.

In this part, reliability will be observed in two following forms for construct interpretation. The first form is the reliability index. The concept is that items produced

out of the same spec and expected to test the same set of competence domains would have comparable task content, a test quality called unidimensionality. When the items are unidimensional, they should likely obtain similar responses in the same testing event from the same examinees. Accordingly, the responses correlate highly with one another and constitute a scale with high test reliability.

The second form of reliability is alphas-if-items-deleted. The concept is that in a unidimensional test, again, the item responses of all the items would vary comparably, and their variances would contribute fairly equally to the reliability index. When one of the test questions is removed, the reliability index of the remaining items should stay relatively stable, which can be particularly evident in a scale-level evaluation after each and every item is taken away in turn. Alphas-if-items-deleted are brought in because of two flaws of the reliability index. A first flaw is that a test of high reliability may not always be unidimensional. A multidimensional test can also achieve a high reliability if the scale variance is high. The second flaw is that multidimensionality of a scale may be hidden under the reliability measure, which is produced as a single number (cf. Cortina 1993: 101f. for details). On these accounts, alphas-if-items-deleted are expected to complement the reliability index in this prototyping. Considering the two statistics gives higher confidence in evaluating viability of the item type in measuring a distinct construct than using a single statistical measure.

In summary, the primary inquiry into the possibility of ConCloze measuring a construct is to examine empirical comparability of a set of unidimensional items. The inquiry aims to find out (a) overall consistency, the reliability of the entire test, and (b) local consistency, the alphas-if-items-deleted. It is hypothesized that the ConCloze items have a high reliability index and consistent alphas-if-items-deleted. It is worth stating that part of this section was presented in Kongsuwannakul (2015a), and the limitation of investigating the construct competence this way will be discussed in Part 5.2 (page 276).

In addition to the main line of inquiry, issues such as appropriateness in the level of test difficulty will also be examined. Examining side issues can be useful because the difficulty of a test created may or may not match the actual performance of examinees (Nissan & Schedl 2012: 292). In the case of mismatch, the test spec can then be adjusted in later research stages so as to be better in line with the average level of the examinees' proficiency.

3.2.2 Test specification (spec)

An idea presented in the previous section is that items of similar test-task content should elicit similar responses in the same testing event from the same examinees. Given this idea, the current spec engineers the test such that the items are as comparable as possible. Also known as test design, the spec is provided below and will be followed by explanations: a sample item in Figure 18 (answer: *A alternative*) and their guiding language in Table 6. All the test items generated are provided in Appendix 1 (pages 290ff.). It is worth emphasizing that upon learning more about the construct competence, the spec will be adjusted and discussed subsequently.

1 on political settlement. TONY BLAIR, BRITISH PRIME MINISTER: There is no _____ but to start from two fixed points of principle in
 2 a California accommodation for smokers is not available as a[n] _____ in cold-weather Chicago. # In fairness to the estimated 60 million Americans
 3 off our coast. But when you think about it, the _____ is not to have a coastline there, because if we
 4 line, and the ability to use a remote control. Another _____ is to use low-cost e-mail-only services, which some local phone
 5 educate him than to kill him, which is the only _____ that has ever been discovered" (Keesing, 190). # Born only fifteen
 6 willingness to imagine both a past and a future political _____ to a degrading but transformable present. And he begins each
 7 to video. Unfortunately, the results were not spectacular. As a[n] _____, Jerry has suggested that local videographers record instruction tapes directly

All the lines above miss the same word. Which of the following should be that word?

A alternative	B choosing	C offer	D possibility
---------------	------------	---------	---------------

Figure 18 A ConCloze sample item

Table 6 Guiding language for ConCloze 1

Entry	Guiding Language (Test Design)
1	There is one sample item (Item 0) and test instructions given at the beginning of the test.
2	Each item has four options, only one of which is the correct answer (key). Selecting one correct answer is assumed to be a task type familiar to the examinees.
3	Selecting the correct answer of each item is scored 1. Selecting any of the other options is scored 0. Not selecting any option is scored 0. Selecting more than one option in each item is scored 0.
4	The prompt is made up of seven concordance lines, each marked with its line number at the front. This emphasizes the fact that they are from different texts.
5	The concordance lines are KWIC-centered and truncated.
6	Each concordance line arbitrarily contains ten words on either side of the KWIC.
7	There is no modification made to the words in the concordance lines. The only exception is when a giveaway of the correct answer would pose a construct-irrelevant threat. In the sample item (Figure 18), the modification is adding square brackets to the article <i>an</i> in Lines 2 and 7.
8	Concordance lines are sampled quasi-randomly from Davies' (2008–) Academic Genre of the Corpus of Contemporary American English (COCA) (cf. page 46 for reasons supporting the genre use). Lines which contain a definition or a definition-like texts of the target words are excluded.
9	All the concordance lines are sorted right to the KWIC blanks in alphabetical order.
10	All the KWIC blanks are fixed at an equal length. In Figure 18, this is three underscores long.
11	The stem wording of each item is constant. It states the problem, "All the lines above miss the <u>same</u> word," and urges action, "Which of the following should be that word?"
12	Each target word is sampled purposely from the somewhat arbitrarily set 1–1.3K range of Gardner & Davies's (2014) Academic Vocabulary List (AVL). AVL is generated based on COCA's Academic Genre, and 1–1.3K means the 1,000 th to 1,300 th most frequently used words in this subcorpus (cf. pages 82ff. for reasons supporting use of this word list; page 236 for an example of ramification in selecting this word list).
13	The distractors are drawn from a close semantic field. This design is driven by the framework for construct measurement (pages 51f.). The distractors can have either a collocational, analytical, or paradigmatic relationship to the target word, in no particular pattern. In Figure 18, <i>choosing</i> has an analytical relationship with <i>alternative</i> , <i>offer</i> a collocational relationship, and <i>possibility</i> a paradigmatic relationship.
14	To find semantically related distractors, the definition of the target word in question is looked up in a dictionary. Then words (a) from the definition or (b) related to the content words in the definition are to be selected purposely as candidate distractors. Alternatively, such related words can be selected from http://wordassociations.net/ .
15	The forms of the distractors must be changed so as to be identical to that of the target word. In Figure 18, all the distractors (B–D) are nouns or equivalent.
16	All the options are checked against the concordance lines and, when deemed necessary, added alternative suffixes to. This must be done such that no testwiseness can give away the correct answer. For example, if the adjective form of Option D <i>possible</i> were used, then test-task engagement might be driven purely by knowledge of derivational morphemes.
17	All the options are arranged in alphabetical order.
18	Options are drawn from three word classes: noun, verb, and adjective. Each has 13 items, arranged in order: noun, verb, adjective, and so forth.

In this study, a key feature in test design is to address possible construct-irrelevant threats (page 65). For example, ConCloze is premised to be a new item format; some examinees may be unfamiliar with the nature of a concordance prompt. They may be unaware that, for example, the concordance lines are retrieved from different texts—an unawareness that may cause their performance to be poor due to the unfamiliarity, rather than because of lacking the proficiency tested. Such unfamiliarity is dealt with in Entry 1 of Table 6 above by providing the examinees with a sample item (Item 0). If the sample item were not provided in the test, the unfamiliarity might cause construct-irrelevant variance and undermine test-score interpretations. Figure 19 illustrates the sample item together with some explanations for the examinees.

Below is one sample test item. You may need to maximize the web browser display to fit the item.

Sample item

1 a close consanguineous relationship. SUBSYSTEMS The comprehensive or master processes _____ all or some of the elements within the social system
 2 staff who should be instructed by practical demonstration how to _____ the alarm system. Every member of the warehouse staff should
 3 with the increasing demand for 90's sets. Thermo-Fixing Spray is _____ ed by heat, either from a hair dryer or heated rollers,
 4 protozoa. It was found that the aerated lagoon was an _____ ed sludge system rather than an oxidation pond. At no time
 5 then whenever that guard becomes ready, either copy may be _____ ed, the choice being invisible to the environment. The two guarded
 6 system tended to break down during the war, but was re _____ ed; it had reached the pre-war level of efficiency by 1951.
 7 anyone at home"? Tanner wisecracked. "Of course", replied Helva logically, _____ ing her outside scanners. "Are you my partner"? she asked hopefully,

All the lines above miss the same word. Which of the following should be that word?

- ☐ A accumulat(e)
☐ B activat(e)
☐ C install
☐ D start

As shown in the sample test item, the question text (=the item prompt) has 7 lines. These lines are from different places, but they need the same word to fill in their blanks. Only one of the four choices can complete them all.

In this sample item, Choice B (activate) is the correct answer, so you choose it by ticking the circle before it.

On the next pages, you will find 39 multiple-choice items of the same format. To get an accurate result, you may wish to imagine that you were taking a final exam, where you had up to one hour to finish these items.

Prev

Next

Figure 19 ConCloze 1 sample item (Item 0)

In tandem with tackling construct-irrelevant variance, the spec writing is also construct-oriented. For example, Messick's (1994) construct-centered approach is applied to item sampling, in which the intrinsic value of a competence domain involved should be the starting point of a validation program. This is realized in Entry 12 of Table 6 above, where target words are selected from a middle-frequency range (1–1.3K) of

Gardner & Davies's (2014) Academic Vocabulary List (AVL). Focusing on the 1–1.3K frequency range is based on two following reasons.

The first reason is its potential pedagogical value. Vocabulary, particularly when its frequency is in the middle of a general word list built from texts of various disciplines, is known to be a most problematic to learners (reviewed in Thurstun & Candlin 1998: 268). The vocabulary is believed to be academic vocabulary (*ibid.*). Accordingly, use of this frequency range in AVL is presumed to likely be reflective of what truly matters to the examinees. On the one hand, any range of the entire AVL could potentially be useful to the learners depending on situations and contexts of language use. Yet, the pedagogical value of this middle-frequency range may also fit the pedagogical purpose of ConCloze testing (page 22). For example, proficient learners may know most words in the 1–1.3K range of AVL when compared with those of a low-proficiency level, who may not know many of the words. Accordingly, the words in the range could function as discriminators between those with high proficiency and those with low proficiency.

In light of using 1–1.3K words as proficiency-level discriminators, a contrastive example is also worth considering. The words that occur very frequently in a corpus (e.g., 0–0.3K, the 300 most frequently-occurring words) might be known even to those of a low general-proficiency level. If they were used as the target words in this test version, a limited language-pedagogical utility of those very frequently occurring words might then be implied for proficiency estimation (*cf.* pages 44ff. for the use of individual words in estimating construct proficiency). This could be so because it would make little sense to administer a language test—except for achievement-testing purposes—only to find almost all of the test takers obtaining full scores on the scale. On this account, selecting this 1–1.3K frequency range may thus fit the ConCloze purpose as a proficiency test, in which examinees of high and low proficiency levels must be separable (*cf.*, for example, pages 22 and 44 for individual words as part of the language sampled to estimate how proficient an examinee would be in language use). The target words (keys) are listed in Table 7 in item order.

Table 7 ConCloze 1 target words (keys)

Item	Target Word	AVL (nth)	PoS*	Item	Target Word	AVL (nth)	PoS
1	<i>coordination</i>	1021	n	21	<i>recreational</i>	1139	a
2	<i>couple</i>	1004†	v	22	<i>advisor</i>	1176	n
3	<i>desired</i>	1007	a	23	<i>structure</i>	1183	v
4	<i>endeavor</i>	1052	n	24	<i>prevailing</i>	1154	a
5	<i>revise</i>	1031	v	25	<i>proximity</i>	1197	n
6	<i>marital</i>	1030	a	26	<i>intensify</i>	1221	v
7	<i>vulnerability</i>	1073	n	27	<i>educated</i>	1174	a
8	<i>mediate</i>	1056	v	28	<i>livestock</i>	1216	n
9	<i>absent</i>	1042	a	29	<i>posit</i>	1258	v
10	<i>fertility</i>	1092	n	30	<i>gradual</i>	1187	a
11	<i>elicit</i>	1083	v	31	<i>petroleum</i>	1234	n
12	<i>applicable</i>	1066	a	32	<i>span</i>	1290	v
13	<i>viewpoint</i>	1110	n	33	<i>plausible</i>	1212	a
14	<i>hypothesize</i>	1084	v	34	<i>succession</i>	1266	n
15	<i>adaptive</i>	1093	a	35	<i>contradict</i>	1315	v
16	<i>academic</i>	1123	n	36	<i>traumatic</i>	1241	a
17	<i>categorize</i>	1124	v	37	<i>elimination</i>	1281	n
18	<i>insufficient</i>	1114	a	38	<i>group</i>	1339†	v
19	<i>monopoly</i>	1150	n	39	<i>privileged</i>	1259	a
20	<i>term</i>	1142	v	Total: 13 items for each PoS			

* Part of speech: n = noun; v = verb; and a = adjective

† Highest frequency used: 1,004th in AVL; Lowest: 1,339th

In addition to its pedagogical usefulness, the second reason supporting a sampling of the 1–1.3K range of AVL is for making construct inferences. The entire AVL ranges between 0–3K, consisting of 3,015 words. By contrast, the frequency levels of the target words range between 1–1.3K, rather than being picked from across all the AVL frequency bands available. On the one hand, the range may be deemed relatively clustered, meaning a limited generalizability of score interpretations to the whole universe of admissible observations (cf. Messick 1989; Kane *et al.* 1999 for generalizability and extrapolability of test scores). Yet, it also shows larger-scale and focused word sampling in this 1–1.3K frequency band. This could then be regarded as precision-oriented and securely grounded in measuring this particular range. Such practice is also known as drawing on a narrow bandwidth with high content homogeneity (cf. John & Benet-Martínez 2000: 352f. for content representation in a measurement scale). On this account, an argument is that ConCloze 1 is reasonably comprehensive and robust in terms of intensity-related fidelity to the construct domain, and accords with the unidimensional design intended (cf. Mowbray *et al.* 2003 for fidelity criteria). As a trade-off over extensive generalizability, a reasonably accurate extrapolation to the target domain may then be claimed for the

current spec (cf. also pages 275f. for a limited power of generalization on the research level).

Apart from the pedagogical usefulness and inferential advantage, AVL itself gives corpus empiricism to the spec. AVL is corpus-derived, and using it offers a systematic approach to sampling words for testing proficiency in language use: the words tested are traceable in terms of corpus frequency. A contrastive example of an approach is Butler's (1991) concordance-based test, whose word selections relied mostly on intuition. This suggests that the current spec seeks transparency in the source of the target words used, in order that future research could replicate this study.

Given that AVL has been selected as the word list for the target words, some other existing word lists are also worth contrasting with it. For discussion purposes, they are listed in Table 8 below in no particular order and will be touched upon in turn. In an overall picture, Coxhead's (2000) Academic Word List (AWL), Paquot's (2010) Academic Keyword List (AKL), and Simpson-Vlach & Ellis's (2010) Academic Formulas List (AFL) are all based on corpora of less than five million words. This is a much smaller corpus size than approximately 120 million words of Davies' (2008–) COCA, which Gardner & Davies's (2014) AVL is based on. An implication is that AVL is very likely to be more representative of how academic language is actually used than the other lists. While a well-balanced sampling of texts may be able to offset the disadvantage of a small corpus to some extent, for the general purpose of word listing without a specific context of use (e.g., a university course module), it seems that the bigger is almost always the better. Moreover, also shown in Table 8, COCA covers over two decades of the text range that could be considered contemporary. This seems to be the broadest in the table for the contemporary period, suggesting that AVL is likely to reflect the current academic language use most closely among the lists discussed. On these accounts of contemporary representativeness, AVL seems to be a suitable choice for applying to selecting target words for ConCloze testing.

Table 8 AVL in contrast with some other word lists

Word List	Source Corpus	Corpus size (million words)	Text range	Wordlist Size	Organization	Key Criteria for Inclusion
Gardner & Davies's (2014) AVL	COCA	120 in Academic Genre (425 in COCA)	1990–2011	3,015 types – 1,991 word families	Frequency-ranked	<ol style="list-style-type: none"> 1. Occurring 50% more frequently than normally expected 2. Well distributed in the disciplines 3. Not a technical word
Coxhead's (2000) Academic Word List (AWL)	Purposely built Academic Corpus	3.5	Mostly from 1993–96, but also from 1961	3,110 types – 570 word families	10 ranked sublists	<ol style="list-style-type: none"> 1. Excluding the most frequent 2,000 words of West's (1953) General Service List 2. ≥ 100 occurrences per word family 3. ≥ 25 in each discipline
Paquot's (2010) Academic Keyword List (AKL)	Professional Corpus, and Student Writing Corpus, consisting of partial combinations of four existing corpora	3	From before 1986 up to 1996	930 lemmas	Grouped by grammatical categories and in alphabetical order	<ol style="list-style-type: none"> 1. Including high-frequency words 2. Keyness, with semantically related words that occur frequently included 3. Well distributed in the subcorpora
Simpson-Vlach & Ellis's (2010) Academic Formulas List (AFL)	Academic speech and writing corpora, consisting of partial combinations of four existing corpora	4.2	From before 1994 up to 2004	207 core formulas, 979 spoken formulas, and 712 written formulas	By discourse-pragmatic functions	<ol style="list-style-type: none"> 1. 3-, 4-, and 5-word sequences with at least 10 tokens per million words, and occurring statistically frequently in the academic genre 2. Distributed well in most of the subcorpora 3. High Mutual Information scores

Apart from the overall evaluative judgments, examining some aspects of the word lists individually also yields similar results. First of all, AWL was created on a basis of mutual exclusiveness with West's (1953) General Service List (GSL), a compilation criterion that could be problematic for word selection (Paquot 2010: 15). The reason is that AWL may fail to incorporate frequently occurring and useful academic words simply because the words have been included in GSL. In fact, language changes through time, but GSL is old and consequently has been criticized for its limited utility (as in including *vessel*, but excluding *computer*; cf. Paquot 2010: 10–11 for further details). Accordingly, word inclusion in and exclusion from AWL may likewise be deemed questionable in terms of coverage and utility; there could be words that have gained popularity over recent years but are not included in AWL, and vice versa (*ibid.*). Since this study aims for generalizability of a score interpretation as well as for task innovativeness (Parts 1.1 and 1.2, respectively), AWL may not be a suitable word list for selecting its target words.

The next word list in Table 8 (page 77) is AKL, which additionally incorporates 2,000 most frequently occurring non-academic words from its specially built corpus. While the inclusion could benefit vocabulary learning, the list might be incompatible with the ConCloze test purpose, which entails academic English (page 22). Moreover, AKL is grouped by grammatical categories and arranged in alphabetical order, rather than by word frequency. This makes the list difficult to be used for manipulating ConCloze-task content based on which word association is likely to be known or unknown to the learners – a frequency-oriented notion in language acquisition underlying the theory of lexical priming (discussed as a genesis of the construct proficiency, pages 54ff.). For these reasons, AKL does not seem to be a suitable option of a word list for ConCloze target words.

The last word list in Table 8 (page 77) is AFL, which contains frequently occurring 3-, 4-, and 5-word sequences. If the formulas in the list were used in place of individual target words in ConCloze items, then the framework for construct measurement could have become highly complicated. For example, if a 5-word sequence were the target formula of a ConCloze item, item writing might require undue care and thus research resources. All the concordance lines are designed to have an equal number of words on either side (ten words in ConCloze 1; cf. Entry 6 of Table 6, page 71), and

so formulas of the same length would be required for the distractors, an extra step in generating items. This is in order to avoid a giveaway by means of comparing the number of words in the options. In fact, before multi-word sequences can serve as target units in ConCloze tasks, a rational step would be to first explore single words for the KWIC position. A final reason is that formulaic language is still little understood at present (e.g., Schmitt 2010: 235), and it can come across as non-contiguous, flexible expressions with slots (Schmitt & Carter 2004: 6f.). For example, *made it clear that*, *making it abundantly clear that*, and *makes it quite clear that* all belong to an identical yet variable formulaic expression (Schmitt 2005–6: 25–6). This implies that operating a formulaic ConCloze could be complicated and involve a language domain that is not yet much understood. Accordingly, AFL would not yet be a suitable word list for the present study (cf. Section 5.3.1 for a recommendation for future ConCloze studies pertaining to formulaic sequences).

In sum, while it could be insightful to seek word/phrase sampling from all the lists discussed, doing so could otherwise demand excessive resources and thus seems infeasible in this research context. Although AVL is relatively new and has not had much research into its validity and properties, it starts to gain momentum in the field. For example, Newman (2016: 33) found that it can better represent core academic words than AWL. For the reasons stated thus far, Gardner & Davies's (2014) AVL appears to be a most suitable choice of a word list for the target words in this study, when compared with the other lists.

In addition to a principled selection of the target words, the item distractors are also engineered strategically. One idea is from the theoretical framework for construct measurement (Part 2.3, page 60), hypothesizing that the ConCloze competence may be acquired along with the extent of interweaving of the mental lexicon (cf. page 64 for a note on *word information*; cf. Nation 2001: 27 for a contemporary classification of knowing about a word). The other idea is that during task engagement, the core components in meaning of the target word would be processed (page 51). Considering these ideas, it thus follows that knowing the information about a target word exhaustively is likely to involve using the lexical network in the prompt for differentiating it from the distractors. For example, in the sample item (Figure 18, page 70), the target word *alternative* denotes one of the choices to pick, and so does the distractor *possibility*. The target word and this distractor are hence in a paradigmatic relationship and share several

semantic components, e.g., *something to pick, one of several*. Differentiating between the two would then require a robust mental lexicon as the examinees need to tease out the target word by extracting word information related to it out of the prompt. Examples of clues for *alternative* include the collocate *available* in Line 2 and a use of the sentence-adverbial position *As an alternative* in Line 7.

In light of the test design for distractor selection, it is worth stating that because the test purpose is to serve as a proficiency test, not a vocabulary test (pages 22 and 44), the frequency level of the distractors is not focused on in the current ConCloze version. This is so because the task content is meant to sample the examinees' performance in the way that whether they can glean and use the lexical network in each item to match with a target word. A proficient examinee in the construct is theoretically expected to be able to do so in a larger number of ConCloze tasks than an inept one. Accordingly, whether or not the examinees know the meaning of each and every word in the options is not a focus for task engagement.

Another example for using word information in the prompt in distinguishing options of semantic relatedness is in Figure 20 below. The key is Option A *coordination*, meaning organizing different elements to work together. The distractor *work* conveys part of this denotation and so is in an analytic relationship with *coordination*. The examinees' ability involved could be to recognize the clues in the co-texts, which would hint at the target word and help them to differentiate the target word from the distractor. Examples of clues may include the collocate *policy* in Line 4 and the co-occurring prepositional phrases *among federal, territorial, and native planning initiatives* in Line 1, *among themselves* in Line 2, and *with care assessment agencies* in Line 6, for instance. Given the concept outlined thus far, the distractors are designed to be selected based on their semantic features in relation to those of the target words, as specified in Entry 13 of Table 6 (page 71).

Item 1

Exit this survey

13%

1 balanced development, emphasizing tourism along the Alaska Highway corridor, and _____ among federal, territorial, and native planning initiatives. Although the results
 2 _____ to international organizations have initiated the practice of consultation and _____ among themselves, often without waiting for specific instructions from their
 3 _____ in Table 1 enabled each group to identify areas for _____ and joint ventures. The development of initial funding mechanisms through
 4 _____ in all of the other military-to-military contacts cited here, policy _____ has not always gone smoothly, and each side has attempted
 5 _____ must be related not only to case management, advocacy, and _____ of agency programs, but also to resource development in particular
 6 _____ for patients over 75 years old should be conducted, improved _____ with care assessment agencies is needed, and community services need
 7 _____ the teacher has the most concrete, practical ability to create _____ s, combinations, and mergers that are the basis for any idea

All the lines above miss the same word. Which of the following should be that word?

- ☐ A coordination
- ☐ B integration
- ☐ C organization
- ☐ D work

Figure 20 ConCloze 1 Item 1

“[W]ords entered into somewhat different semantic relations according to their part of speech” (Read 1993, referenced in Read 2012: 312). Given this, the spec also aims to randomize the parts of speech of the options. Dealt with in Entry 18 of Table 6 (page 71), words of three main word classes of content words—noun, verb, and adjective—are sampled. As listed in Table 7 (page 75), the words are then arranged with systematic interspersions of one another throughout the test. Randomizing the parts of speech is expected to balance out the types of lexical network that may be invoked, insofar as the item responses could be interpreted for construct proficiency with relatively higher confidence. For example, if the target words in this prototyping phase were all adjectives, then the operationality on the scale level might be limited chiefly to the networks of a predicate after a linking verb and of a noun after an attributive adjective – the grammatical structures that might eventually support the dimension of grammatical use in the test purpose rather than also that of vocabulary use (cf. the test purpose on page 22).

AVL originates in Davies’ (2008–) COCA, one of the largest corpora available presently (Schmitt 2010: 312). Given this, a first point to stress in the test spec is genre specificity, which has been discussed in terms of language representativeness in Chapter 2 (page 46) and is dealt with by the spec in Entry 7 of Table 6 (page 71). Moreover, taking concordance lines exclusively from COCA could create general consistency in test-task content and spelling and hence can reduce construct-irrelevant variance from different spelling systems and word use, for example. Following Fulcher’s (2003b: 135ff.) concept of fixed elements for task content in spec writing, this practice could be deemed desirable for test design and will therefore be held constant throughout this study.

To clarify the point of genre specificity, an opposite scenario is also worth considering. Figure 18 (page 70) illustrates the possibility of concordance lines being deliberately sourced from multiple genres, viz. Lines 1 and 3 sourced from COCA’s Spoken Genre, Line 2 News, Line 4 Magazine, and Lines 5, 6 and 7 Academic (cf. Kongsuwannakul 2015b for malleability of the ConCloze item type). This study presumes that words may behave variedly according to their genres (cf., e.g., Flowerdew 2000; Bhatia 2008; Hancioğlu *et al.* 2008 for similar insights). Mixing concordance lines from different genres may thus produce a confounding factor in language variation, which would make it hard to control task content. Therefore, for minimal construct-irrelevant variance, this spec fixes COCA’s Academic Genre as the only source of concordance lines (cf. also pages 46 for the genre considered in light of the theoretical framework for

construct measurement). An item example made exclusively from this genre is in Figure 20 above.

Given that the source of concordance lines has been selected, a distinction in assessing language ability is also worth reiterating. Even though the target words and concordance lines are all derived from Davies' (2008–) COCA, this does not mean that academic vocabulary knowledge, for example, would be the hypothetical construct domain (cf. Section 1.3.2 for the hypothetical construct). On the one hand, proficiency in academic English grammatical and vocabulary use may be the current hypothetical construct (pages 22 and 46), which may thus encompass knowing the meaning of words used frequently in the academic genre. On the other hand, however, the target words and eventually the concordance prompts function as probing tools for estimating how proficient an examinee would be in dealing with many different pieces of test-task content (page 44). This thus means that the proficiency estimation is based on the total likelihood of integratedly mobilizing such knowledge domains as lexical-semantics to complete a series of ConCloze-test tasks meaningfully and consistently throughout a testing (cf. pages 40f. for an important role of the norm-referenced assessment paradigm in realizing a meaningful estimation).

Concerning the size of the concordance prompt, the spec takes common practices into consideration, the first of which is about familiarity with texts. For example, Stevens (1991b) created concordance-based exercises using a corpus specially built from his students' own course books. Given this, the exercises could be assumed to contain only texts familiar to the students. Because his concordance for each KWIC was only three–four lines long, and the exercises were implemented successfully, their content must have been adequate for task completion. An inference is that three–four concordance lines may be minimally sufficient when familiarity with the subject matter and genre can be assumed. Accordingly, the number of concordance lines for ConCloze should increase in the absence of such familiarity.

Another consideration for an appropriate concordance size is about truncating concordance lines. Butler (1991) asserted that approximately four complete sentences were needed for his concordance-based cloze test (see also Part 1.2 about his test). An inference is that four concordance lines may be needed when a set of concordance lines are not truncated and familiarity with the concordance texts cannot be assumed. ConCloze

features a truncated concordance prompt of unseen texts; therefore, its number of concordance lines presented to the examinees should be greater than four. Considering the practices in both Stevens's (1991b) and Butler's (1991) concordance-based tests, over four concordance lines seem to be needed for ConCloze. Given that the ConCloze design features truncated concordance lines, seven concordance lines are decided, albeit somewhat arbitrarily, as a preliminary format feature in Entry 4 in Table 6 (page 71). If empirical evidence is found to indicate that the decision is unlikely to be appropriate for the examinees' meaningful task engagement, changes can be made to the feature in later ConCloze versions.

In addition to setting the number of concordance lines for the prototype, the spec also seeks to determine an appropriate size of the concordance lines. A first consideration is about working memory, which can (a) hold information for a short while before forgetting, or (b) turn to misattribution—incorrect memory because of, for instance, reactivity of close semantic associates presented sequentially (both reviewed in Schacter 1999). The capacity of the working memory for a correct immediate recall is 7 ± 2 information chunks (i.e., pronounceable, two-second-long texts and numbers each) (Shiffrin & Nosofsky 1994). Assuming one information chunk is equivalent to approximately one–two words, ten words could probably be an arbitrary but likely suitable size for each concordance line that can be expected not to impose an onerous cognitive burden on examinees and thus construct-irrelevant variance on test interpretation. This design is described in Entry 6 of the spec (Table 6, page 71) and is also illustrated in Figure 20 (page 81).

Given the determination of the line size that may operate efficiently, a second consideration is also worth taking into account. Corpus-retrieved sentences can vary substantially in length, ranging from those a few words long to the sentences that may take several lines on a normal A4-size paper. Accordingly, using a full sentence retrieved from COCA for each concordance line could otherwise turn the size of a concordance prompt out of control. For example, rather than exactly 70 words for every one of the item prompts across the test, some prompts may be in the region of 40 words whereas the others exceed one hundred. Unless needed for a purposeful task-content manipulation, this would then leave the overall picture of task content in chaos. Is it, for instance, ample clues about the key KWIC or the semantic network among the options that affects the examinees' success rates of task completion? On this account, fixing the number of words

for each concordance line could be deemed a practical option for test design as it systematizes the production of the test items and renders the responses to them comparable for validity interpretations.

3.2.3 Population and sample

ConCloze is generated using key words and concordances from an academic genre. Given this, the intended population is defined as non-native speakers of English who engage or have engaged with academic English. In operational terms, they include, but are not limited to, those studying in or having graduated from the university level, because they are likely to have exposure to academic English (cf., e.g., Graddol 1997; Schneider 1997; Crystal 1997, 2003; Jenkins 2003 for a central role of English in higher education). This definition of the intended population is applicable throughout this study.

For practicality, convenience and snowball samplings are used in this part and for the rest of this thesis. The sample consists of those studying in a university program or having graduated from higher education (undergraduate and postgraduate). Their data are collected through an online platform called SurveyMonkey, whose limitations together with those of the sampling methods will be discussed in Part 5.2 (page 281). The test webpage is <https://www.surveymonkey.com/s/AcadEnglishVocabTest>, open for one month. 38 participants are recorded as visiting the webpage and starting the test log. Figure 21 illustrates the number of participants throughout the test administration.

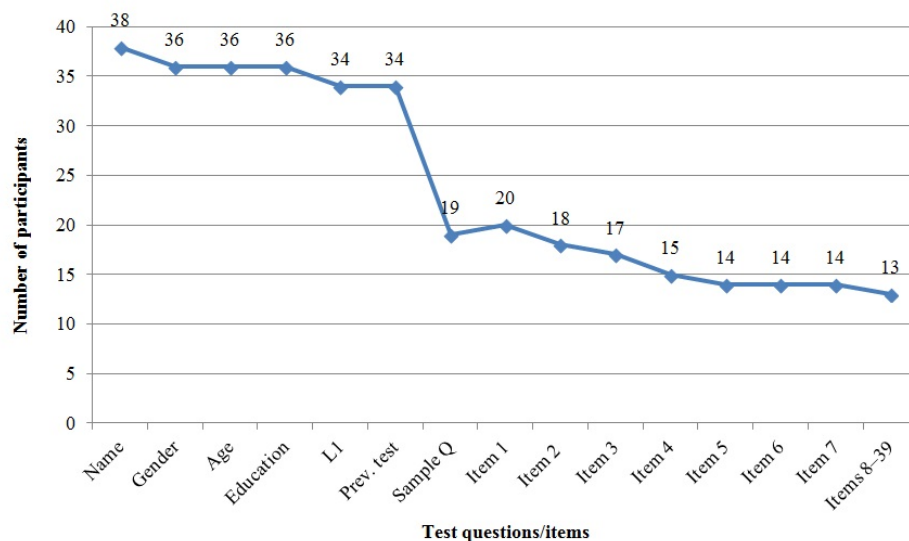


Figure 21 Participant number of ConCloze 1

This research follows BERA's (2011) guidelines of research ethics: the examinees can opt out anytime if they wish. Depicted in Figure 21 above, 13 of the initial number 38 completed the entire test, which accounts for a 34% completion rate and a dropout rate at 66%. Figure 21 also shows a sharp decline in the number of examinees when the sample item is presented to them, and the number then becomes constant by the end of Item 8. It may be inferred that once the respondents explored the sample question, many of them did not wish to continue the testing. After trying answering some early test questions, a few more respondents also decided to exit the test webpage. According to Nissan & Schedl (2012), trying out a prototype can adequately use as few as ten respondents. Therefore, 13 test completers collected might be deemed generally adequate for initial investigations of the current prototyping phase.

Given the anonymity of participating in the test, those who opted out before test completion cannot be contacted for their reasons for dropping out. Nonetheless, analyzing the test-completers' comments (to be detailed on pages 100ff.) reveals that test difficulty could be an issue: the difficulty of ConCloze 1 is found to potentially exceed the average level of the examinees' language ability. An inference about the sampling is that the low completion rate could partly be accounted for by an unsuitable level of test difficulty. It is worth stating that finding an unreasonably high level of test difficulty can be problematic, especially for proficiency tests like ConCloze. This is so because a low scorer, for example, may in fact have a moderate level of the construct proficiency but

obtain a low score because the test is too difficult for him/her. Accordingly, the low scorer cannot be separated from those who really have low proficiency, and the test may be considered failing to fulfill its intended purpose.

The sample can be described as follows. In Table 9 below, the gender distribution is shown, in which the majority of the respondents (58%) are female. It can be inferred that the sample's item responses are distributed reasonably equally between the genders. Then, the age of the respondents is analyzed in Table 10 below, in which their ages range from 19 to 53 (range=34), and the median is 33. An inference is that the responses of the sample are distributed broadly across age ranges. Afterwards, their education levels are displayed in Table 11 below, wherein the majority of the respondents (77%) are associated with the postgraduate level. This implies that the sample likely belongs to the population defined in this study. Based on these descriptions, it might be unlikely that the item responses in this part would be severely biased towards one particular category of these variables. Also, an argument is that the convenience and snowball sampling methods might be reasonably appropriate for collecting initial evidence of the task performance reflective of the intended population's construct competence.

Table 9 Gender of ConCloze 1 respondents

Gender	Count	Percent
Male	14	38.9
Female	21	58.3
Prefer not to answer	1	2.8
<i>Total</i>	36	100

Table 10 Age of ConCloze 1 respondents

Age	Number	Count	Percent
31, 39	3 each	6	17
22, 26, 28, 30, 33, 35, 37, 38	2 each	16	44
19, 23, 25, 27, 29, 32, 34, 36, 40, 41, 42, 46, 48, 53	1 each	14	39
<i>Total</i>		36	100

Table 11 Education levels of ConCloze 1 respondents

Highest Education Level*	Count	Percent
Preseasonal course to an undergraduate level	4	11.43
Year 1, undergraduate	1	2.86
Year 3, undergraduate	2	5.71
Holds a bachelor's degree	1	2.86
Preseasonal course to a postgraduate level	12	34.29
Year 1, taught postgraduate	2	5.71
Year 1, research postgraduate	3	8.57
Year 4, research postgraduate	2	5.71
Year 5, research postgraduate	1	2.86
Holds a master's or a Ph.D., or studies at an year-unspecified postgraduate level	7	20.00
<i>Total</i>	35	100

* Zero-response and N/A categories are excluded from presentation

In addition to the sample's gender, age and education level, responses on their language profile are also collected. The first variable is L1. Displayed in Table 12 below, those speaking Thai as their L1 are 22 in number (65%), and those speaking Arabic seven (20%). An inference could be that the data might not be well distributed in terms of L1 diversity. Then Table 13 below shows the respondents' experience in living or staying in English-speaking countries, a variable intended as a surrogate for their intensive exposure to English (see also Adolphs & Durow 2004 for sociocultural integration as a qualitative factor). Twenty (59%) have stayed no more than three months, and seven (21%) between three–six months. It could be interpreted that the majority of the respondents might not have had much of such exposure.

Table 12 L1s of ConCloze 1 respondents

L1	Count	Percent
Arabic	7	20
Kazakh	2	6
Kurdish	3	9
Thai	22	65
<i>Total</i>	34	100

Table 13 Length of staying or living in English-speaking countries

Length	Count	Percent
0–3 Months	20	59
3–6 Months	7	21
6–9 Months	1	3
9–12 Months	2	6
12 Months–2 Years	1	3
2–5 Years	1	3
Over 5 Years	2	6
Total	34	100

The last variable in the sample's language profile is the result of a latest standardized English test. Shown in Table 14 below, nine (28%) had never taken any of such tests. Up to the period of data collection, 14 (42%) had taken at least one test in the past two years. Overall, the range among those who reported taking a test spans over ten years (before 2004 to 2014). It is worth stating that, collectively, 17 of the respondents (52%) had taken an IELTS test during the past three years. Among them, 13 also reported their overall results, the mean of which is 5.54 on Bands 0–9. An initial implication could be that the respondents' previous tests of English could vary widely temporally, and they may tend not to have taken an identical test. The descriptions thus far mean it could be difficult to compare target respondents validly based purely on their length of intensive exposure (discussed earlier) and standardized English-test results.

Table 14 Year of latest standardized English test

Year	Count	Percent
2014	6	18
2013	8	24
2012	3	9
2008	1	3
2007	1	3
2004	1	3
Before 2004	1	3
Unspecified	3	9
N/A	9	28
Total	33	100

3.2.4 Test responses

The test spec has been provided in Section 3.2.2, and the population and sample defined in Section 3.2.3. In this section, raw scores by test completers will be first

explored for initial evidence of a competence operating during ConCloze-task engagement. The score range is between 10 and 28 of the maximum possible score of 39. The average is 18, and the standard deviation of the scores 7.04. This average is equivalent to 46% of the maximum, with a wide score dispersion of 14 (\bar{x} 18 \pm SD 7.04). An interpretation of these test statistics is that the sample appeared to perform slightly lower than the midpoint of the scale (19.5), suggesting that the test was somewhat difficult for them. Yet, the average score and the score dispersion may indicate that the items have probably elicited systematic responses, which is a basic quality of most psychometric measures. Suggesting a trait underlying the examinees' responses, an implication is thus that the ConCloze item type might be used for measuring a distinct construct competence.

The test items in this part are designed to be unidimensional (page 67). Given this, the wide score dispersion could be explained as follows. The items are deliberately engineered to be comparable in terms of task content (page 69). Therefore, the examinees could have responded to them reasonably consistently—either consistently high or consistently low. This allows an inference that those who scored well consistently might be more proficient in mobilizing the construct competence measured by ConCloze (cf. page 44 for a description of a score–proficiency relationship). By contrast, those scoring consistently low could be less proficient in so doing. On this account, ConCloze scores might be inferred to have internal consistency, suggesting a first piece of initial evidence of the existence of a language competence potentially underlying task performance in this item type.

In addition to the basic test statistics, test reliability is also examined. The index is 0.84, which indicates high test reliability. Obtaining a high reliability index suggests that each examinee deals more or less consistently with the test tasks throughout the test. This evidence accords with the earlier finding on the likelihood of systematic item responses, signifying that the items may test the same domain of competence, and hence adds weight to the inference that the scale may be viable in measuring a distinct construct.

Considering the pioneering nature of ConCloze 1, the reliability coefficient may seem unusually high. This can be explained by two reasons, the first of which is the unidimensionality of the scale already anticipated (page 67). The second reason is test length: Yang *et al.* (2004) developed a multidimensional scale and claimed that seven items is sufficient for seeking internal consistency in each dimension. ConCloze 1 is

unidimensionally engineered and has 39 test items, approximately five times the appropriate size contended by Yang *et al.* (2004). Thus, the high reliability could also be attributed partly to an inflation effect of the test length.

In tandem with the reliability index, alphas-if-items-deleted are also investigated. Table 15 below shows the statistics in item order; for example, deleting Item 27 can increase the reliability of the scale to 0.85 at best. The alphas are found to vary from 0.82–0.85 in the deletions. Because the reliability of the entire test is 0.84 (discussed earlier), an interpretation can be that deleting any particular item does not seem to make much change to the whole-scale reliability. Rather; each item is likely to contribute fairly equally to the measure.

Table 15 Alphas-if-items-deleted in ConCloze 1

Item	Alpha if item deleted	Item	Alpha if item deleted	Item	Alpha if item deleted
1	0.83	14	0.83	27	0.85*
2	0.84	15	0.84	28	0.83
3	0.85	16	0.85	29	0.82
4	0.83	17	0.84	30	0.84
5	0.82†	18	0.85	31	0.85
6	0.84	19	0.83	32	0.83
7	0.82†	20	0.84	33	0.84
8	0.84	21	0.84	34	0.83
9	0.84	22	0.85*	35	0.83
10	0.83	23	0.84	36	0.84
11	0.83	24	0.84	37	0.83
12	0.83	25	0.83	38	0.83
13	0.85	26	0.84	39	0.84

* Highest

† Lowest

High internal consistency can indicate that the items measure the same competence (Instructional Assessment Resources 2011). Given this, examining alphas-if-items-deleted of all the items across the scale highlights the consistency of the examinees in responding to the items. Indicating homogeneous task content, this gives another piece of construct-related evidence: structural validity. When considered in conjunction with the high reliability index (page 90), the evidence suggests that measuring a discrete construct through ConCloze is likely to be viable.

In addition to the internal consistency, patterns in item responses are also explored. They are primarily those of item difficulty (also called item easiness or item

facility, IF) and item discriminability (also called item discrimination, ID), which are fundamental item qualities indicating the viability of construct measurement (cf. Nissan & Schedl 2012: 283, 285). With regard to item difficulty, appropriateness of the difficulty level is usually required to be considered while prototyping an item type. The goal is to ensure that the test is doable and not so difficult for the intended population that the score interpretations would otherwise be invalid (i.e., response invalidity – invalid test interpretation because the examinees cannot engage in the test tasks meaningfully). Determining such appropriateness could be through several sources of information, the first of which can be the stakes of the testing. For example, Nissan & Schedl (2012: 283f.) contended that approximately 80% can be deemed appropriate for a prototype of a TOEFL section, which generally has high stakes. ConCloze could be considered low-stakes because a respondent's performance in this study would not mean, for instance, a huge testing fee for retaking the test or affect the examination results of their university courses enrolled. On this account, it does not seem justified in this study to aim for an average score as low as 50% of the maximum possible score. Therefore, the mean score of ConCloze 1 at 46% could be argued as inappropriately low.

Besides weighing the percent average score against the test stakes, two other sources of information are also evaluated in determining appropriateness of the test difficulty. The first is the sample's education level, which shows the majority of them are from the postgraduate background (page 88). Given the possibility that they would generally have had more years of exposure to academic English than undergraduate respondents, the average score is yet lower than the midpoint (page 90). This suggests that, again, ConCloze 1 would be inappropriately difficult for the intended population, which is non-native speakers of English who engage or have engaged with academic English (page 85). The second source of information used for determining appropriateness of the test difficulty is opinions of the test completers. Discussed extensively in Section 3.2.5 (pages 101f.), it will be demonstrated that the level of test difficulty could be considered inappropriately high. On these accounts, the difficulty level of the current prototype may be regarded as inappropriate.

In addition to the average score, individual items also reflect viability of the scale in measuring a construct. Table 16 below displays their IFs, which are categorized into three groups by traditional cut-off points: 0–0.29 = difficult, 0.3–0.7 = moderately difficult, and 0.71–1 = easy. The IFs are found to range between 0.15–0.85, and 74% of

the items (29 of 39) can be categorized as moderately difficult. Because the moderate range is usually held good (Brown 1996: 76ff.), only a minority of the items can be argued as either too easy or too difficult. For general testing purposes, this could be deemed a good spread of IFs, indicating that the items can potentially be refined and item-banked according to their difficulty. While this potential involves test use rather than test interpretations, it demonstrates usefulness of the item type, which is part of the foundation of validity (cf. Messick 1993 for appropriateness, meaningfulness, and usefulness as inseparable essence for validity).

Table 16 ConCloze 1 item difficulty and discriminability

Item	Item Facility (IF)		Item Discriminability (ID)		
	Item Difficulty	Evaluation [†]	Point Biserial Correlation [‡]	Corrected Item–Total Correlation [‡]	Evaluation
1	0.45 (20)*	Moderate	0.48	0.43	Acceptable
2	0.17 (18)*	Difficult	0.06	0.01	Poor
3	0.76 (17)*	Easy	–0.15	–0.21	Poor
4	0.53 (15)*	Moderate	0.47	0.43	Acceptable
5	0.36 (14)*	Moderate	0.81	0.82	Acceptable
6	0.57 (14)*	Moderate	0.20	0.14	Poor
7	0.36 (14)*	Moderate	0.65	0.79	Acceptable
8	0.23	Difficult	0.26	0.21	Acceptable
9	0.15	Difficult	0.15	0.11	Poor
10	0.85	Easy	0.49	0.46	Acceptable
11	0.39	Moderate	0.54	0.51	Acceptable
12	0.62	Moderate	0.61	0.59	Acceptable
13	0.54	Moderate	–0.24	–0.32	Poor
14	0.62	Moderate	0.7	0.69	Acceptable
15	0.46	Moderate	0.33	0.28	Acceptable
16	0.23	Difficult	–0.10	–0.17	Poor
17	0.31	Moderate	0.24	0.18	Poor
18	0.62	Moderate	–0.09	–0.16	Poor
19	0.39	Moderate	0.58	0.56	Acceptable
20	0.31	Moderate	0.24	0.18	Poor
21	0.62	Moderate	0.41	0.36	Acceptable
22	0.31	Moderate	–0.05	–0.12	Poor
23	0.31	Moderate	0.31	0.26	Acceptable
24	0.39	Moderate	0.20	0.14	Poor
25	0.31	Moderate	0.62	0.6	Acceptable
26	0.46	Moderate	0.11	0.04	Poor
27	0.77	Easy	–0.34	–0.4	Poor
28	0.69	Moderate	0.61	0.6	Acceptable
29	0.31	Moderate	0.81	0.82	Acceptable
30	0.23	Difficult	0.44	0.41	Acceptable
31	0.31	Moderate	0.07	0.01	Poor
32	0.31	Moderate	0.55	0.57	Acceptable
33	0.54	Moderate	0.44	0.4	Acceptable
34	0.54	Moderate	0.66	0.64	Acceptable
35	0.39	Moderate	0.61	0.59	Acceptable
36	0.77	Easy	0.44	0.41	Acceptable
37	0.62	Moderate	0.74	0.74	Acceptable
38	0.31	Moderate	0.5	0.47	Acceptable
39	0.77	Easy	0.29	0.24	Acceptable
Average	0.46		0.35	0.31	

* indicates an actual sample size in calculation different from 13, the number of participants who completed the entire ConCloze 1.

[†] Criteria: 0–0.29 = difficult, 0.3–0.7 = moderately difficult, and 0.71–1 = easy

[‡] Criterion: ≥ 0.2 = acceptable discriminability

An item response is a product of an examinee's competence in interaction with test-task content. Given this, the varied IFs in Table 16 above might indicate the task

content that varies when processed by the examinees. In other words, the present variation in IFs can be viewed as the manifestation of a relatively constant competence tackling test tasks that vary in difficulty. Based on this interpretation, the implications for validity investigation are threefold. First, the task content could contain clues that are crucial to ConCloze processing and that contribute to the difficulty variation. Secondly, assuming the variation of the clues in task content, they could be harnessed by adjusting item components, which may help to determine the substantive content as processed using the underlying competence. Lastly, the present variation in difficulty is the first item-level evidence in this study to suggest viability of the items in measuring a discrete construct. It also exhibits congruence with the scale-level evidence such as the reliability coefficient and alphas-if-items-deleted (pages 90f.).

With regard to patterns of IFs, Table 7 (page 75) shows ConCloze 1 items in descending order of the frequency levels of the target words. For example, the target word of Item 1 *coordination* is the 1,021st word in AVL, *endeavor* of Item 4 1,052nd, and *vulnerability* of Item 7 1,073rd. Generally, language learners are exposed to fewer lower-frequency words than higher-frequency ones (Schmitt 2010). Accordingly, lower-frequency words are less likely known to them than higher-frequency words (Nation & Waring 1997; Shaw & Weir 2007; Richards *et al.* 2008). Considering this likelihood, it would follow that lower-frequency words would make items more difficult than higher-frequency words. However, Table 16 demonstrates that the IFs do not systematically descend. For example, the IF of Item 3 (1,007th) is 0.76 whereas the IFs of Items 36 (1,241st) and 39 (1,259th) are both 0.77. This means that as the items pass and the test proceeds, the frequency-based item ordering may not necessarily result in increased difficulty. Two inferences can be made based on this finding. First, this lack of incremental item difficulty is contrary to the frequency-based ordering. While learners may likely have greater exposure to higher-frequency words than lower-frequency words, the frequency levels of the target words may not affect item difficulty linearly. Also, it is unlikely that the IFs would systematically reflect how the mental lexicon is organized and drawn upon during task engagement. Secondly, it is worth restating that the target words are sampled from a narrow frequency bandwidth (page 75). Hence, the current finding could suggest a result of the sampling, in which the frequency levels of these target words may be so close that no patterns of IFs are observable. In whichever case, an implication could be that ConCloze-item difficulty tends not to vary by the corpus-based frequency

of target words, meaning that the examinees are unlikely to process them only. As to the finding about a lack of discernible frequency-induced IF patterns, a later test version (ConCloze 6, pages 186) will deal with it systematically.

In addition to the IFs, IDs of ConCloze 1 items are also shown in Table 16 (page 94). Power of discrimination is when a test item can separate an examinee who has a higher level of the competence being measured from those who have a lower level. Discriminability can indicate that the items test a particular competence (Jackson *et al.* 2002; Embretson 2007). Depending on the objective of a test and on the sample's ability range, an ID of approximately 0.2–0.25 is generally deemed satisfactorily discriminating (cf. Henning 1987: 53 for flexibility in setting discriminability criteria). ConCloze 1 has low stakes (page 92); therefore, a criterion used in Table 16 (page 94) is arbitrarily set to be that both the point-biserial correlation and the item–total correlation must be greater than 0.2. In the table, the average point-biserial correlation is found to be 0.35, and the average corrected item–total correlation 0.31. Sixty-four percent of the items (25 of 39) pass the criterion, meaning that the majority of the items discriminate quite well. An inference based on these results is that the items, even though first prototyped in this study, seem to test the same domain or set of competence domains systematically. Therefore, the ConCloze item type is likely to be viable for measuring a discrete construct.

In addition to analyzing responses to the keys, responses to the distractors of the items will also be explored. Seeking possible patterns emerging from the responses to all of the options can be useful because it is found earlier that the examinees may not process only the target words during task engagement (page 96). It seems reasonable to hypothesize that the distractors could play a role in task processing. Given this hypothesis, the responses to the items in Table 16 (page 94) that are both moderately difficult and well-discriminating are explored in Table 17.

Table 17 Responses to items of good qualities

Item	Target Word (Key)*	Count	Distractor 1	Count	Distractor 2	Count	Distractor 3	Count
1	<i>coordination</i>	9‡	<i>integration</i>	9‡	<i>organization</i>	1	<i>work</i>	1
4	<i>endeavor</i>	8†	<i>attempt</i>	6†	<i>difficulty</i>	1	<i>fruitlessness</i>	0
5	<i>revise</i>	5‡	<i>arrange</i>	1	<i>change</i>	6‡	<i>consider</i>	2
7	<i>vulnerability</i>	5‡	<i>awareness</i>	6‡	<i>revelation</i>	2	<i>strength</i>	1
11	<i>elicit</i>	5‡	<i>inform</i>	0	<i>question</i>	1	<i>response</i>	7‡
12	<i>applicable</i>	8†	<i>included</i>	0	<i>suitable</i>	5†	<i>true</i>	0
14	<i>hypothesize</i>	8‡	<i>formulate</i>	1	<i>suggest</i>	2‡	<i>verify</i>	2‡
15	<i>adaptive</i>	6‡	<i>evolutionary</i>	2	<i>individual</i>	2	<i>responsive</i>	3‡
19	<i>monopoly</i>	5‡	<i>abuse</i>	1	<i>capitalism</i>	4‡	<i>competition</i>	3
21	<i>recreational</i>	8‡	<i>active</i>	2‡	<i>distracting</i>	1	<i>enjoyable</i>	2‡
23	<i>structure</i>	4†‡	<i>arrange</i>	4†‡	<i>found</i>	1	<i>integrate</i>	4‡
25	<i>proximity</i>	4†‡	<i>distance</i>	3‡	<i>immediacy</i>	3†	<i>region</i>	3‡
28	<i>livestock</i>	9‡	<i>breeding</i>	2‡	<i>creature</i>	1	<i>stray</i>	1
29	<i>posit</i>	4†	<i>argue</i>	3	<i>explain</i>	5†	<i>suggest</i>	1
32	<i>span</i>	4‡	<i>include</i>	1	<i>prolong</i>	5‡	<i>stretch</i>	3
33	<i>plausible</i>	7‡	<i>explanatory</i>	2	<i>ingenuous</i>	3‡	<i>satisfactory</i>	1
34	<i>succession</i>	7‡	<i>appointment</i>	1	<i>empire</i>	2	<i>position</i>	3‡
35	<i>contradict</i>	5‡	<i>confound</i>	4‡	<i>propose</i>	2	<i>refute</i>	2
37	<i>elimination</i>	8†	<i>completion</i>	0	<i>process</i>	2	<i>removal</i>	3†
38	<i>group</i>	4†	<i>divide</i>	2	<i>gather</i>	6†	<i>select</i>	1

* Word class of options*: noun in Items 1, 4, 7, 19, 25, 28, 34, and 37; verb in Items 5, 11, 14, 23, 29, 32, 35, and 38; adjective in Items 12, 15, 21, and 33

† indicates potentially competing responses because of synonymous relationship

‡ indicates potentially competing responses because of other semantic relationships

In Table 17 above, the target word in each item is marked, and so is at least one other distractor which has a competing response count. The response count of a distractor is considered competing when it is the highest among the counts of the distractors. For example, in Item 4, the target word *endeavor* garners eight responses. The distractor *attempt* receives six responses, which is the highest count among those to the distractors. *Endeavor* and *attempt* might be regarded as having similar meanings, so they are both marked with † for their synonymous relationship. Another example is Item 5, the key of which is *revise*. The most attractive distractor is *change*—its potential hypernym—and hence the ‡ marking.

In analyzing the patterns of response counts, it is initially found that responses to the keys and distractors are unlikely to be random. For example, Item 11 obtains 13 responses, which are not distributed equally among the options but accumulate at the distractor *response* and the target word *elicit*. In fact, distractors in 35% of the items (7 of 20) can be systematically marked distinctive with a synonymous relation with the target words, and distractors in 75% of the items (15 of 20) with other semantic relations.

Irrespective of the duplication in Items 23 and 25, the responses to these distractors seem to form a pattern: distractor words which share semantic components with the keys appear to attract substantial responses to themselves. The sharing of the semantic components tends to be either in part or in full. For example, in Item 5, the distractor *change* and the target word *revise* can both refer to amendment. Yet, *change* could be regarded as a more generic word than—and hence a hypernym of—*revise*. Another example is Item 34, where the distractor *position* competes with the target word *succession*, which refers to taking over a position. *Position* also denotes part of the core semantic component, and hence is a meronym of *succession*. This initial discovery of patterning of competitiveness in some distractors seems to indicate that lexical-semantics is likely to be a domain of language knowledge used by the examinees.

Entry 13 of the test spec (Table 6, page 71 and the discussion on page 79) determines that the distractors are semantically related to the keys. This design is intended to investigate if the examinees can meaningfully differentiate the target words from the distractors. In light of the distractor functioning, the design can be argued to have an impact on task processing: distractors may have become nearly correct answers when some of their semantic components are similar to those of the keys. Because synonyms and words of related meanings bear semantic co-referentiality (Crossley *et al.* 2014, in print), an argument could be that the keys and nearly correct distractors may have co-referents, some of the clues of which would be found in the concordance prompt. Given this, deciding between the keys and the distractors based on the concordance-based co-referential clues could be a language process mobilized in this item type. It is worth restating that the hypothetical construct will be used for construct evaluation in light of the test purpose and the construct-related findings thus far in the concluding part of this version (Section 3.2.6).

3.2.5 Textual feedback

In the previous section 3.2.4, test and item responses have been focused on as a source of validity evidence. In this section, respondents' textual feedback will be examined so as to provide backing to the evidence discussed. The feedback is collected immediately at the end of the test (*post hoc* elicitation), meaning that the respondents' data obtained may be deemed an immediate retrospective account of their testing

experience. On the one hand, this feedback may be unable to reflect the experience on the level of individual items. Yet, it can give insights into their summative assessment of the testing. The form used for collecting the feedback is illustrated in Figure 22.

98%

You're great! You have finished the new academic English vocabulary test.

Your comments are always welcome. Please write some feedback about your test experience in the space provided below.

Test instructions (Clear about what to do?)

Sample item (Clear and concise?)

Test length (the 39 multiple-choice items)

Test design (the overall feature, e.g., one page for one item)

Item design (Comprehensible? Doable? Are four choices too easy or too difficult?)

Other comments

Prev Next

Figure 22 ConCloze 1 comment boxes

The questions for collecting the respondents' feedback cover general issues in test and item design. For example, the first question illustrated in Figure 22 is about the clarity of the test instructions, and the second question the clarity of the sample item provided. These questions are designed such that flexible responses could be evoked, which could

help to identify construct-irrelevant issues that may stem from non-task content such as test length. Other issues that could be present yet unanticipated can also be identified with an open-ended question, collected in the last question of Figure 22. The responses are summarized in Table 18 below by the questions and proportions of positive feedback. It is worth stating that while care is exercised in interpreting the feedback, the comments could sometimes be indecisive in terms of polarity (positive/negative).

Table 18 ConCloze 1 respondents' comments

Question	Proportion of Positive Comments (%) [*]	Example (<i>Pseudonym</i>) [†]
Clear test instructions?	12 of 13 (92.31) [‡]	Of course, they are clear. (<i>Holly</i>)
Clear and concise sample item?	12 of 12 (100)	Clear and precise (<i>Gary</i>)
Appropriate test length?	2 of 13 (15.38)	good (<i>Mark</i>)
Appropriate test design?	10 of 12 (83.33)	Easy to read and friendly-user (<i>Amy</i>)
Comprehensible item design?	5 of 13 (38.46)	it is good enough to use the context to pick up the right word. (<i>Isaac</i>)
Other comments?	1 of 8 (12.5)	it was useful but difficult (<i>Mark</i>)
Total	42 of 71 (59.16)	

^{*} The comment page is quasi-optional: the respondents are asked to answer at least three of the questions provided. Accordingly, the percentages vary according to the actual number of the examinees who responded to each question.

[†] No editing performed

[‡] Italicized percentage indicates that the majority of the comments to the question are positive.

Overall, in Table 18 above, 59% of all the comments (42 of 71) are positive. Individually, out of the six questions in the table, half receive from the test completers positive comments as their majority. These questions are on the test instructions, sample item, and test design. On this account, the overall test seems to be fairly satisfactory to the sample. It may be inferred that the non-task content as it stands, particularly in relation to the test instructions, sample item, and test design, is unlikely to be in need of major improvements for subsequent research stages.

Notwithstanding the overall positive feedback, three other questions in Table 18 above receive negative responses as their majority: test length, item design, and 'Other comments', which will be tackled in turn. Regarding test length, shows, the majority of the test-completers' comments (7 of 13, 54%) are shown in Figure 23 below, saying that ConCloze 1 was excessively long. An inference could be that part of the intended population might also find ConCloze 1 inappropriately long. On the one hand, a long test

may not necessarily be problematic as long as a sufficient number of respondents can be sought for the analyses required. Yet, long testing can deter prospective respondents, which may stretch research resources needlessly. Another issue is the possibility of response invalidity: a too long test may not encourage respondents of varied proficiency levels to finish it. This would then lead to a question as to whether the test interpretations can represent a wide range of proficiency levels in the target population.

Test: Test length (the 39 multiple-choice items)*

Amy: Quite long.

Bella: Too many items to finish.

Cara: Too many.

Daniel: Too much, no more 30 items that reasonable.

Elle: Too long for me to read and concentrate.

Fatima: So longer.

Gary: Quite long in length.

Holly: 39 items are too much for evaluating only one skill.

Isaac: I am wondering why 39. 39 could be meaningful to the researcher.

Jason: Too long.

Karl: Ok.

Lana: Very long.

Mark: Good.

* Spelling corrected; pseudonymized

Figure 23 ConCloze 1 respondents' comments on test length

With regard to the question on item design in Table 18 (page 100), 38% of the comments (5 of 13) did not express positive comments. All of the comments are displayed in Figure 24 below, indicating that most of the respondents focused more on the aspect of item difficulty rather than, for example, the number of options. In fact, the respondents are found to also reflect in the 'Other comments' box that the test was too difficult for them. While these opinions are subjective in nature, the chance could be that part of the intended population could likewise view the test as inappropriately difficult. This implies that on a large-scale administration and in subsequent versions of this study, the level of test difficulty should be lowered, insofar as to mitigate response invalidity in the data set.

Test: Item design. Comprehensible? Doable? Are four choices too easy or too difficult?

Amy: Interesting and challenging to find the right answer.

Bella: A bit too difficult.

Cara: Too difficult, not enough clues to guess.

Elle: Too difficult.

Fatima: Too difficult.

Jason: Difficult.

Karl: Difficult.

Test: Other comments?

Amy: It seems to me some extracted sentences are quite academic and too difficult to understand in a brief time.

Bella: The test is quite long and tough.

Daniel: If you're not my friend, I will not do this test until finished.

Elle: Sorry I can't stand answering difficult questions for a long time.

Karl: It was useful but difficult.

Figure 24 ConCloze 1 respondents' comments on item difficulty

3.2.6 Conclusion and decision

In this part, ConCloze item responses have been investigated and found having potentially systematic IFs, IDs, and internal consistency. The initial findings of such systematic score variances would mean that the responses are likely to be governed by a discrete competence that accounts for their variability. Accordingly, a preliminary positive answer may be offered to Research Question 1 in Table 1 (page 16), in which item responses of the prototyped test format are hypothesized to be internally consistent, thereby suggesting structural validity. A purpose of test prototyping is to seek initial evidence of the possibility of the item type measuring a discrete competence, and accordingly seems to be fulfilled by the findings, which are summarized in no particular order in Table 19.

Table 19 Major evidence and inferences

Entry	Evidence	Inference
1	High test reliability	Items testing the same domain of competence
2	Consistent alphas-if-items-deleted	Items testing the same domain of competence
3	Varied IF values	Varied test-task content tackled by a constant competence
4	No pattern between IFs and target words' frequency	Task-content variation not solely dependent on the accessibility of the target words
5	Satisfactory IDs of the majority of items produced	Items testing the same domain of competence
6	Attractive distractors with semantic components shared with the keys	Lexical-semantic knowledge mobilized in test-task engagement

In Chapter 1, the test purpose has been set hypothetically to serve as a proficiency test on professional and academic English grammatical and vocabulary use (page 23). The findings reported in this ConCloze version would allow an initial appraisal of the construct, in that the proficiency in vocabulary use may be involved in task processing. As evidenced initially by some discernible patterns of the distractors across a number of items (page 97), lexical meaning, particularly core semantic components of the option words, may have an important role in task engagement. In Chapter 2, a lexical network in the item prompt has been hypothesized to be formed by the associations of the target word in the task (page 50). On this account, the inference about the lexical semantics may indicate that the lexical network in each item is likely to be used in connecting to the target word. But when a distractor happens to share core semantic elements with the target word, then the connecting may be more difficult and hence result in the attractiveness of that distractor. Accordingly, in light of the construct appraisal thus far, a proficient learner with a high level of the construct competence would be able to differentiate slight differences in semantic elements of words more effectively than a learner with low proficiency. Nonetheless, it is worth pointing out that no distinguishable item responses can yet be observed in this version in support of the aspect of grammatical use.

While the construct-related inference appraised above could only be tentative because of the limitation inherent in the small sample of test prototyping, it seems that the item type might be able to fulfill the purpose of a proficiency test. This is in the sense that lexical networks of the target words across multiple items would mean the degree of how expansive and dense the mental lexicon an examinee has in dealing with the test tasks. For example, when the mental lexicon of an examinee is fully developed, the chance is that the majority of the associations in the item prompt would be recognizable as related to the target word and hence lead to a successful task completion. Likewise, the examinee would be able to repeat the success in many more items in the rest of the test, and accordingly could be considered to be proficient in vocabulary use. That said, there has not emerged evidence to support its grammatical aspect of the hypothetical construct yet.

Despite the finding that ConCloze item responses might be explained by a distinct competence, a few issues emerge that should be addressed in subsequent test versions. The first issue is that the respondents may not be well diversified in terms of their L1s,

which could be dealt with by logistic measures such as deliberate diversification of respondents' backgrounds. Although convenience and snowball samplings are used in this study, and hence L1 background is not a decisive factor in screening the participants, diversifying the L1s of the respondents would increase the power of generalization of the finding to the universe of admissible observations, thereby benefiting the inference making.

Apart from the issue of L1 diversity that subsequent test versions should attempt to address, a second issue requiring a systematic solution is on the completion rate. In this version, the rate is low, potentially attributed to inappropriate item difficulty and test length, which may in turn suggest test inertia, a source of response invalidity to test interpretations (cf. Henning 1987: 91ff. for sources of response invalidity). Changes recommended to tackle this issue could be shortening the test and varying task content. This is in order that item difficulty can be reduced as theoretically expected. While such changes may originate in item statistics, a central notion in assessing language proficiency is whether a test interpretation will be valid for a decision to make (O'Loughlin 2011). As a purpose of ConCloze testing is to evaluate if a respondent, albeit anonymous or pseudonymized, could be deemed proficient in the construct being defined (pages 42ff.), being able to do so precisely towards the construct proficiency, rather than under the external influences such as test inertia, is of immense importance. On this account, efforts in subsequent test versions should be put into making the test relatively more accessible by, for example, shortening it and varying its task content.

A third issue calling for a systematic tackling is about being unable to compare the respondents' language profiles for construct interpretations. The majority of the respondents do not have much intensive exposure to English. Nor are their results of standardized English tests validly comparable. This issue implies that the sample's general proficiency in English could not be compared systematically with ConCloze performance. A solution could be seeking data on English tests or equivalent that would put the respondents on the same ground for association with ConCloze. In light of these issues, a decision is that the test development should proceed for test improvement and more evidence for construct interpretations.

3.3 ConCloze 2–4: Qualitative Prototyping

3.3.1 Rationale

In the previous part, quantitative responses have been focused on in prototyping a ConCloze format. For rigorous pre-operational testing, a qualitative prototyping is usually also recommended (cf. Kenyon & MacGregor 2012; Nissan & Schedl 2012 for a need to use both approaches). Incorporating a qualitative prototyping can be useful for three reasons. First, this study uses internet-based testing (page 85). A qualitative prototyping allows checking the mechanisms of test delivery such as clickability of the buttons, which may otherwise hinder optimal performance if malfunctioning. With a robust measure used, the item responses elicited are unlikely to be tainted by usability problems. Secondly, a qualitative investigation also allows real-time observations of task engagement, which can lead to an improved testing experience for the examinees—a cost- and time-efficient approach to developing ConCloze before a larger-scale administration.

Lastly, another usefulness of qualitative prototyping is to investigate substantive processing underlying responses to ConCloze items. In the quantitative ConCloze 1, it is found that the respondents may not process merely the key words, and the semantic components of distractors may have a role to play during task engagement (pages 96ff.). However, a possibility of extreme scenarios still persists. For example, the item responses might in fact represent systematic guessing, in which the respondents did not read the concordance prompt but picked options based on item numbers or some random words found in the test. Investigating the substantive processing can therefore check if the underlying process(es) is comparable from person to person. The idea is that if the tasks draw on comparable processes of multiple examinees, then the item responses can be deemed evidence for the same domain of proficiency (Nunan 1991). In tandem with ConCloze 1, the current prototyping makes a triangulated effort to assess the viability of the item type in construct measurement, which can then increase confidence in evaluating the potential of the item type in measuring a distinct construct.

Accordingly, for a robust delivery mechanism, improved testing experience, and substantive-validity evidence, ConCloze will be prototyped qualitatively in this part. This begins in Section 3.3.2 with adjusting the old test spec to the current testing. Then in Section 3.3.3, usability issues of the online test platform are explored, in which the respondents are hypothesized not to struggle with the online testing. Afterwards, verbal

reports are analyzed in Section 3.3.4, where common processes and task content verbalized during task engagement will be focused on for one test item. Considering the findings in ConCloze 1, it is hypothesized that the task content of (a) the concordance prompt, (b) options, and (c) meaning of option words are commonly verbalized by the respondents. Finally, construct inferences and decision based on empirical evidence in this part will be summarized in Section 3.3.5.

3.3.2 Test spec

In ConCloze 1, items are produced out of a single spec. For example, item components such as number of options and number of concordance lines remain the same throughout the test (cf. page 71 for details). Given the lack of variation in item components, they may be called *univariant*—representing a single item variation only. In this part, an idea is to create multiple item variants (IVs) based on those items. Obtaining item responses through multiple IVs can be useful because intervariant consistency indicates a core construct cutting through amidst variation, another facet of structural-validity evidence. When the current findings are considered alongside those from the univariant items in ConCloze 1, confidence in prototyping ConCloze as a new item type can increase, and the validity argument could be sounder as a result.

Prototyping is a test-of-concept exploration, which does not necessarily involve a very large sample size (Petre & Rugg 2010: 90). Given that multiple IVs are to be created, five items are selected from ConCloze 1 and turned into IVs. The items are provided in Appendix 2. Their spec is presented in Table 20, and explanations will follow thereafter.

Table 20 Guiding language for ConCloze 2

Entry	Guiding Language (Test Design)		
1	ConCloze 1 presumes that a sample item is necessary for task engagement (page 71). In order to verify this, no sample item is provided at the beginning of ConCloze 2.		
2	At the beginning of the test, the test instructions, “Choose the most appropriate answer” is provided.		
3	In between the item pages, there is one ‘no-task’ page. This is in order to prevent the respondents from continuing without any pause and being interfered by the next task. The page also provides room for eliciting their immediate-retrospective accounts, if any or deemed insightful, of their task performance.		
4	Modifications made to the five ConCloze 1 items are:		
	ConCloze 1 Item* (Original IF)	Modification Applied	ConCloze 2 Item
	4 (0.53)	N/A	1
	12 (0.62)	Three options	2
	14 (0.62)	Five words on either side of the KWIC	3
	21 (0.62)	Semantically unrelated or distantly related options	4
	28 (0.69)	ten concordance lines	5


* For example, no item component of Item 4 from ConCloze 1 is modified. Only the item number is changed into Item 1 here.

For parsimony, Table 20 draws only distinctions between the ConCloze 1 spec and the current spec. A significant modification is its Entry 1. It is presumed in ConCloze 1 that concordance-based testing is specialist-niched, in that only those who have had hands-on experience of corpus querying will know of concordances (page 72; cf. also Sinclair 2004a; Kilgarriff 2009 for a comparatively exclusive position of applied corpus linguistics for language teaching). Providing a sample item has back then been vital to comprehending the test-task content, particularly that contained in the concordance prompt. In this part, the presumption is going to be verified by withdrawing the sample item. This is intended such that some form of miscomprehension could be detected in the verbalizations if concordance-based testing is really specialist-niched.

Regarding number of IVs, Fulcher (2003a: 395) recommended that in one test administration, no more than eight item prototypes be presented to the examinees. Given this, five is arbitrarily determined for the number of IVs in Entry 4 (Table 20 above), which is within the recommended limit and is unlikely to cause much confusion to the respondents, an otherwise construct-irrelevant variance. The precursor items from ConCloze 1 are chosen stratified-randomly from the moderately difficult and

satisfactorily discriminating items in Table 17 (page 97). This selection approach is used because their original item qualities (base difficulty and discriminability) should still remain in part when the modifications are introduced. Figure 25 illustrates an example of an IV, which features three options (answer: *A applicable*).

Exit this survey



University of
Leicester

Testing a Qualitative Aspect of Academic English Vocabulary (Prototype Version)

The second question

44%

1

traits, which are characterized by relative stability over time and _____ across situations, positive state-like capacities are relatively more malleable and

2

urbanization display three research perspectives: abstract frameworks that _____ are universally _____ but at the expense of substantive specificity; context-dependent frameworks that

3

require the use of resources that require you clear the _____ copyright laws before viewing and/or listening to them with your

4

aim for a break, just for socialist reform, a reform _____ in the new areas opened by the revolution and its

5

still to be resolved. (n374) "Of 188 HAPs, 134 have _____ measurement methods. Forty-three have likely methods. Eleven have potential or

6

efficient and biologically appropriate. While this solution may not be _____ to all common property resources, for those cases which may

7

Court held that so long as a law is generally _____, and not discriminatory in being specifically aimed at religion, the

All the lines above miss the same word. Which of the following should be that word?

☒ A applicable

☐ B suitable

☐ C true

Prev

Next

Figure 25 A ConCloze 2–4 Item 2

Number of options is a known factor affecting discriminability of item distractors as well as task validity (Rodriguez 2005; Baghaei & Amrahi 2011; Lee & Winke 2013). Given this, one of the item distractors is deliberately eliminated in Entry 4 of Table 20 (page 107). The elimination takes effect in Item 2 (introduced earlier, illustrated in Figure 25 above). The aim is to examine if options are really verbalized by ConCloze respondents as inferred in ConCloze 1 (pages 97f.). Administering both three-option and four-option IVs alongside can be useful because their operability will suggest congruence in test-task content as processed by the same respondents and impart intervariant reliability—i.e., internal consistency of the multivariant test planned in the rationale (pages 105f.). This line of reasoning applies similarly to the other IVs in Table 20.

In addition to the item spec, an *ad hoc* measure is also adopted in test administration. Given the prototyping nature, the researcher must be present, and dynamically observe the respondents' behavior while they engage with the online test. When prompted or required by their apparent struggle regarding usability issues, the researcher may intervene, explain to, or facilitate the respondents. This is in order that the verbalization is as smooth as possible and is thus least affected by construct-irrelevant threats, if any. Tackling such struggle will then inform how to improve the spec and testing experience in a later version. It is worth stating that the current qualitative prototyping is rapid-iterative: minor changes to the specs that follow will also be presented later.


3.3.3 Testing usability

In testing usability of a computer-based prototype, Fulcher (2003a) stated that there can be several interface–design issues identified and addressed. This may range from hardware specifications to software ones, such as font display, color scheme, and page navigation. According to Fulcher, such testing can be performed in a process called rapid iteration: a small cohort of participants each gives feedback, and the problems will be identified and fixed accordingly. Then the next cohort does the revised test and gives feedback on it. Also, the respondents are usually observed whilst engaging in the test task, and then give accounts of the testing experience through, for example, questionnaires or interviews (cf. Bachman & Palmer 1996; Nissan & Schedl 2012). In case of giving think-

alouds, one cohort may comprise approximately 4 ± 1 respondents, from whom at least 75% of the usability issues should be detected (Nielsen 1994).

One of the current objectives is to investigate the usability of computer-based, multiple-choice ConCloze. The line of inquiry is to probe whether the testing is likely to be suitable for the intended population. The focus is on its functionality because the testing may or may not require special training or modifications in order for the respondents to be able to sit it. Given this, there are two primary aspects of usability considered here: test presentation (page navigation and clickability of the webpage elements displayed) and task specification (test instructions and clarity of the task), which will be dealt with in turn. Along the discussion, other peripheral issues will also be addressed in order that the testing experience is optimized. It is also worth restating that the topics of (a) whether the sample item is clear and concise, (b) whether the number of options (four) is appropriate and doable, and (c) whether the test length as it stands is appropriate have all been covered in ConCloze 1. Therefore, these topics will not be investigated in this section.

The procedure of usability testing can be divided into four major steps. First, an online sign-up form is created (cf. a sample page in Figure 26 below; available then at <https://www.surveymonkey.com/s/InvitationToAcademicVocabTest>). An invitation message to it is sent out and forwarded electronically to prospective respondents.



**University of
Leicester**

Sign-up for Testing Academic English Vocabulary (Prototype Version)

2. Biodata

40%

This page asks about your details such as gender and age.

You are

☐ Male
☐ Female
☒ Prefer not to answer

How old are you?

Age

25
 16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Can I have your email address, please? (Check the spelling carefully, as this is for contact purposes.)

kk234@le.ac.uk

What program are you in?

☐ Preessional course to an undergraduate level
☐ Preessional course to a postgraduate level
☐ Year 1, undergraduate
☐ Year 2, undergraduate
☐ Year 3, undergraduate
☐ Year 4, undergraduate
☐ Year 1, taught postgraduate
☐ Year 2, taught postgraduate
☐ Year 3, taught postgraduate
☐ Year 1, research postgraduate
☐ Year 2, research postgraduate
☒ Year 3, research postgraduate
☐ Year 4, research postgraduate
☐ Year 5, research postgraduate
☐ Other (please specify)

Prev

Next

Figure 26 A ConCloze sign-up page, with representative examples

The second step is to screen for potential participants. When prospective respondents have filled out the sign-up form and given consent to audio recording, some of them are selected according to the diversity in demographic information they can contribute to the study. Illustrated in Table 21 below, part of this information is listed individually. For example, the age of those selected ranges between 20 and 46, and educational backgrounds from undergraduate to postgraduate research-based levels. Of importance is their language profile, in which their L1s range from Cantonese to Thai (nine L1s), and their IELTS scores from 5.5 to 8.5. Diversifying the respondents'

backgrounds deliberately implies that the task responses elicited could likely represent the population better than those from, for example, a totally homogeneous sample.

Table 21 Demographic profile of ConCloze 2–4 respondents

Name*	Age	Level†	L1	Latest Standardized Test Reported	
				Year	Result
ConCloze 2					
Aaron	27	PGR	Chinese	2010	IELTS: 6.5
Björn	22	UG	Cantonese	2013	IELTS: 6.5
Claire	36	PGR	Korean	2010	IELTS: 6
Dakota	30	PGR	Chinese	2010	IELTS: 6
Esther	35	PG	Urdu	2008	IELTS: 8
ConCloze 3					
Franz	33	PGR	Thai	N/A	IELTS: 7
Gill	46	PGR	Dagbani	1997	IELTS: 7.5
Halle	20	UG	Romanian	2011	IELTS: 8.5
Igor	36	PGR	Thai	2010	IELTS: 5.5
ConCloze 4					
James	27	PGR	Sinhala	N/A	N/A
Klavier	20	UG	Hungarian	2012	IELTS: 6.5
Lulu	35	PGR	Chinese	2007	IELTS: 8

* All pseudonymized

† UG = undergraduate; PG = taught postgraduate; PGR = research-based postgraduate

The third step in usability testing is eliciting verbal reports, in which the respondents are first invited individually for a verbalization. Prior to the beginning of each session, the link to the test webpage is sent to the participant (ConCloze 2 available then at <https://www.surveymonkey.com/s/LgProcessPrototype>; ConCloze 3 <https://www.surveymonkey.com/s/testquestion>; and ConCloze 4 <https://www.surveymonkey.com/s/NewTestQuestion>). Sending the test link only before the verbalization session is intended to replicate the individual testing experience, to the extent that the clickability of the test link can be examined. During the testing, the respondent is observed whilst clicking the test link and starting doing the test. Notes are made when their struggle with the interface features can be observed or when they express problems going through the test. The last step in usability testing is when a next ConCloze version is improved using the information obtained: information from ConCloze 2 is used for improving ConCloze 3, and information from ConCloze 3 for ConCloze 4. In total, there are 12 respondents for the usability testing, equaling 12 verbalization sessions with 60 verbal reports (12 respondents \times 5 items). The adequacy of this sample size will be justified empirically considering (a) a proper functionality of the test platform and task

format and specification (page 123), and (b) the processing saturation identified on-line (page 128).


In investigating test presentation, general observations are made on, for example, whether the respondents can easily navigate from one page to another. They are also observed when selecting an option and filling out the textboxes, for instance. Basically, (a) the clickability of the link, buttons and icons, and (b) the functionality of the checkboxes and of the textboxes provided are checked. For illustrative purposes, Figure 25 (page 108) is worth considering, in which a ConCloze item with three checkboxes and two buttons is illustrated. The checkboxes represent the three options of the item, and the buttons ('Prev' and 'Next') are for page navigation. Then in Figure 26 (page 111), a page of the online sign-up form is depicted, showing some representative examples of four questions and their answers: gender (checkbox), age (drop-down list), email address (textbox), and educational program (checkbox).

Whilst eliciting verbal reports, it is observed that none of the respondents had difficulty in clicking the test links (for ConCloze 2–4 each) or navigating across pages. Nor did they struggle with selecting the options and filling out the textboxes. The textboxes are used for starting the test log with a name entry and for collecting the end-of-test feedback. Observing these aspects during task performance allows an inference that the buttons, checkboxes, and textboxes are unlikely to pose usability issues affecting the respondents' test-task completion. Thus, a construct-irrelevant threat from page navigation and webpage functionality to the construct interpretation is unlikely.

Two possibilities could explain the lack of discernible problems in test presentation. First, the test platform is on the commercial SurveyMonkey.com, which provides tested templates for hosting the ConCloze test in general and for its construction, user-friendliness and maintenance in particular. Another explanation is that the links to the sign-up form and the tests are all sent out electronically, suggesting that basic computer and internet literacy may be assumed *a priori*. Therefore, it may be concluded that the overall test functions well in terms of presentation, and the ConCloze platform is unlikely to require any special training of the intended population to be able to take the online test.

In light of the functionality of the delivery mechanism, the usability testing will proceed with investigating task specification. First of all, each respondent is observed in

such a way as to whether they can do and complete the task, irrespective of whether their answer is right or wrong. Because the primary task is to read the concordance prompt and select one of the options, the respondents' understanding of the test instructions ('Choose the most appropriate answer.') and their ability to choose an option are here focused on. For illustrative purposes, Figure 27 shows the test instructions, which are located over the first item.



Testing a Qualitative Aspect of Academic English Vocabulary (Prototype Version)

Now, the academic-vocabulary question prototypes.

31%

You may need to maximize the web browser display to fit the question.

Choose the most appropriate answer.

1 theological detail ultimately undermines the properly interdisciplinary, nature of the _____ and results in a dominance of the social sciences in
2 their world, but that education in such a primary human _____ as music should be universal for our children in their
3 _____ logical to go with your strengths and to structure the _____ in such a way as to guarantee success by anticipating
4 and coordination across divisions of the college must occur. This _____ is a campuswide effort that begins with a coordinating committee
5 or adventure travel. It is a tedious and time-consuming _____ that requires considerable inconsequential travel over vast expanses of territory.
6 devoted to social activity would leave less time for academic _____s and therefore would likely have a negative impact on academic
7 being answered, but, as is always the case in scientific _____s, more questions have arisen. We haven't yet found evidence for

All the lines above miss the same word. Which of the following should be that word?

☐ A attempt
☐ B difficulty(/ie)
☐ C endeavor
☐ D fruitlessness(e)

Prev

Next

Figure 27 ConCloze 2–4 Item 1

Regarding understanding test instructions, it is found that none of the respondents expressed their incomprehension of the test instructions or the question stem ('All the lines above miss the same word. Which of the following should be that word?'). Nor did they inquire into the meaning of particular words used therein. Nonetheless, there is some variation in how the respondents expressed their understanding. Exemplified in Figure 28 below, some respondents started engaging in the test task immediately, without

verbalizing the test instructions or the question stem altogether (as in Case 1). In other cases, they explicitly verbalized these components of task specification and showed their understanding of what to do (as in Case 2). Some others started the task and quickly returned to asking for confirmation of what they believed they were expected to do before proceeding with the task (as in Case 3). It is worth stating that because all the ConCloze 2–4 items have a comparable format, Figure 28 presents only some variations discovered in Item 1.

Case 1

Researcher: OK, just say whatever you are thinking.

Gill: OK, erm er [Line 1] ‘theoretical detail [‘ultimately’ not verbalized] undermines the properly interdisciplinary nature of [E] the [E] [KWIC position here] [P] and results in a dominance of the social sciences’ [P] er [IA] [IA Option D ‘fruitlessness’] [P]

Researcher: Just keep saying.

Case 2

Researcher: Whatever you are thinking, whatever you are er what is going on in your mind, just say it.

Igor: OK, I’m [Item instructions] ‘Choose the most appropriate answer.’ [P] appropriate answer, then I will have to choose only one, right?

Researcher: Yeah.

Igor: From er four choice.

Researcher: Yes.

Igor: The first one [concordance line] is er

Case 3

Björn: Question is t [Line 1] ‘theoretical details ultimately undermines the properly interdisciplinary [P] nature of the erm [KWIC position here]’ [P] [IA ‘what’] I so put [P] in between them

Researcher: Yes.

Björn: Er [Line 1 continued] ‘[KWIC position here] and results in a dominance [P] of the’

Figure 28 Comprehending the test task

All the verbal reports contain verbalizations of the concordance lines and of some or all of the options, and similarly culminate in an option being selected. Given this, the variation as exemplified in Figure 28 is unlikely to demonstrate the respondents’ incomprehension of the test task. Rather, it indicates their situational decision to verbalize or not to verbalize this content of the test. An inference is that the test instructions and question stem are sufficiently clear and thus are likely to be comprehensible to the target population. This means that the test instructions and question stem should not be major sources of construct-irrelevant variance to test interpretation.

With regard to choosing an option, it is observed that none of the respondents had difficulty in ticking the checkbox in order to choose an option. Nor did they specify in their feedback that the very action of choosing an option was a difficulty to them. Further, again, all of the verbal reports end with selecting an item option, which amounts to a decision made. This can be deemed a sign that the respondents had likely engaged in the test task. Assuming that the multiple-choice task format is widely used (Parshall *et al.* 2002), the three pieces of evidence gathered imply the respondents' ability to choose an option in the current multiple-choice ConCloze format. They potentially (a) had read the test instructions and/or the question stem and understood the test task well, or (b) took it as a matter of course to choose one of the options provided. As it stands, the task of option selection seems appropriate for the intended population. Accordingly, the evidence thus far generally indicates adequacy and appropriateness of the task specification; this usability aspect is unlikely to require any extensive modifications for the population to take the ConCloze test.

Closely related to the respondents' understanding of the task specification is their understanding of the prompt. In ConCloze 2, 8% of the verbal reports (2 of 25 [5 respondents \times 5 items]) are found to contain traces suggesting the respondents might not fully understand that a concordance is made of lines retrieved from different places, the very characteristic of the concordance. Figure 29 illustrates the occurrences.

Case 1

Björn: What [IA]. All the questions is from same erm same article?

Researcher: No.

B: No? OK.

R: They are from different places.

B: Different places. Hmm. [P] what's what's [IA]

Case 2

R: Say it.

Claire: [Line 3 continued] 'by anticipating' [Line 4] 'and coordination across divisions'

R: These lines are from different places. Uh-huh. They are from different places. But they miss the same word.

C: [...]

R: These lines are not connected. Alright? They are from different places. They are not from the same passage.

C: [P] Hmm [P] But, you know, it's the, the first sentence [i.e., Line 1] it ends on the number five [i.e., Line 5]

R: No, they are not connected; these sentences are from different places. One to seven [i.e., Lines 1–7] are from different places.

C: [P]

Figure 29 Potential incomprehension of the nature of the concordance

An example showing that the nature of the concordance prompt is not well understood is Case 1 in Figure 29 above. Respondent Björn—22 of age, undergraduate—seemed to express doubt when inquiring if the lines were from the same text. This incident might be interpreted in two different ways along the level of seriousness it may pose to the validity investigation. A first explanation is that the respondent did observe the entire item format. This possibility takes into account the distinctive format of the concordance prompt, and the discrete numbering of concordance lines (cf. Figure 27, page 115). Björn might know that the lines were not to be read connectedly; his question rather functioned as a mere doubt seeking confirmation of his understanding. On this account, the question might at most indicate his attempt to connect the messages or themes that the different lines offered—an attempt which could then become infeasible to him. The other interpretation is that he failed to observe the item format and the alignment of the concordance lines. The expression of uncertainty suggests his attempt to read all the lines interconnectedly as a single running text. This second interpretation could be considered a very similar process to Case 2 respondent Claire's in Figure 29.

The latter interpretation above could be relatively worrisome for validity investigation vis-à-vis the former one. It implies that some respondents' organizational perception towards the concordance prompt as a unique structure might be inaccurate. On the one hand, the fundamental task of choosing one of the options was clear to them and accomplished throughout. Yet, some of their task engagements might have been plagued by their interpretive inaccuracy: a distorted interpretation of the nature of the concordance. In whichever case, if such an incident happens proportionately in the item responses of the population, then the construct-irrelevant threat towards task performance could be significant and undermine test-score interpretation. As such, if there is no correction or modification to the test, such a distortion perceived may persist in subsequent ConCloze versions.


In light of the traces of incomprehension, it could be said that concordance-based testing is not readily comprehensible to the entire target population based solely on test instructions. Instead, its underlying premise that ConCloze is an innovative item format (page 4) is supported. On this account, Entry 1 in the guiding language of ConCloze 2 (Table 20, page 107) does not gain adequate support for continued use. It is decided in this stage (i.e., observation of ConCloze 2 responses for ConCloze 3 preparation) that

there should be a re-introduction of a sample item to the beginning of the test. Hence, a revision to the guiding language for ConCloze 3ff. can be described in Table 22.

Table 22 A revision to the guiding language of ConCloze 3ff.

Entry	Guiding Language
1	Given an empirical finding that some respondents may not be totally familiar with the nature of the concordance, a sample item and an accompanying explanation are to be provided at the beginning of the test.

Determined by the spec revision, a sample item with explanatory notes is created and placed at the beginning of the test. Depicted in Figure 30 below, it is a simplified modification of Figure 19 (page 73), albeit not reflecting the iterative revisions that are going on with the test. Providing the sample item is expected to introduce item components, clarify the nature of the concordance prompt, and make the test task more readily comprehensible to the target population. Therefore, a recommendation for all the subsequent ConCloze versions is that a sample item should be given before the actual tasks. This could mitigate construct-irrelevant variance from an incomprehension of the item format.



University of
Leicester

Testing a Qualitative Aspect of Academic English Vocabulary (Prototype Version)

Sample item

42%

You may need to maximize the web browser display to fit the question.

The multiple-choice question below is a sample item. Study it carefully.

Note that Lines 1-7 below are texts taken from different sources, so they are not of the same passage. However, they need the same word to fill in their blanks.

1 potential cheaters, those who wish to ____ early and those who may need to use the

2 A working model that is not yet ____ ed but that represents the major technical,

3 will discuss it with them after they ____ ed. Do NOT tell them that you will be

4 each lexical item. || Once you have ____ ed the interview, compare the results from

5 A, and the curve for the non-masters ____ es at point C. There will be some non-masters

6 and receive a score directly after ____ ing the test. || It is probably noticeable that

7 deal of important information || To ____ this section, I will present an example of what

Special text, made of excerpts from different sources.

Lines 1-7 here are not from the same place.

All the lines above miss the same word. Which of the following should be that word? ← question

A allow

B finish

C participat(e)

D study(i)

← choices

Prev

Next

Figure 30 ConCloze 3 sample item

In addition to the nature of the concordance prompt, another peripheral issue identified is font size. Collected at the end of the test, the respondents' feedback on this issue can be summarized in Figure 31 below, where 40% of the ConCloze 2 respondents (2 of 5) thought that the font size used in some parts of the test was too small.

Test: Please give comments on font size (Large enough? Easy to read?)*.

ConCloze 2

Aaron: Yes too many words in one question makes me feel stressed

Björn: Bigger one will be better

Claire: Something parts are small

Dakota: Ok

Esther: Yes

ConCloze 3

Franz: Large enough

Gill: Yes

Halle: It is large enough.

Igor: Yes. It is easy to read.

ConCloze 4

James: Good

Klavier: Yes

Lulu: Appropriate

*Spelling not corrected. First letter capitalized.

Figure 31 Respondents' comments on font size

Given the potential problem of font size to some ConCloze 2 respondents, a generic solution is to improve the appearance of words on screen. On the one hand, the appearance is partly dependent on the configurations of individual web browsers, meaning that the font size cannot be completely controlled at the server's end of the test administration. Nonetheless, it is determined that the appearance of the concordance prompt presented should be the priority in the current improvement as it is presumably an area of intensive processing for task completion. On this account, there are two modifications performed locally then to improve this aspect of testing experience.

The first modification is a revamp of the resolution and size of all the concordance prompts used. This is done by increasing the size of .jpg appearance of the prompt before transferring to the test-creating template. It is considered here a way of reducing the effects of differing settings in different web browsers. This modification takes place in the corrections for ConCloze 3 onwards.

The second modification is an addition to the test directions. Exemplified in Figure 32 below, it reads that the respondents can use the zoom-in function ('Or alternatively, you may hold 'Ctrl' and press '+' or '-' to zoom in or zoom out, respectively.'). This can increase the size of the general display as well as of the concordance prompt. The addition is at the beginning of the test over the sample item, as

well as over the main test instructions and the first item. The modification takes place in the corrections for ConCloze 4 onwards.

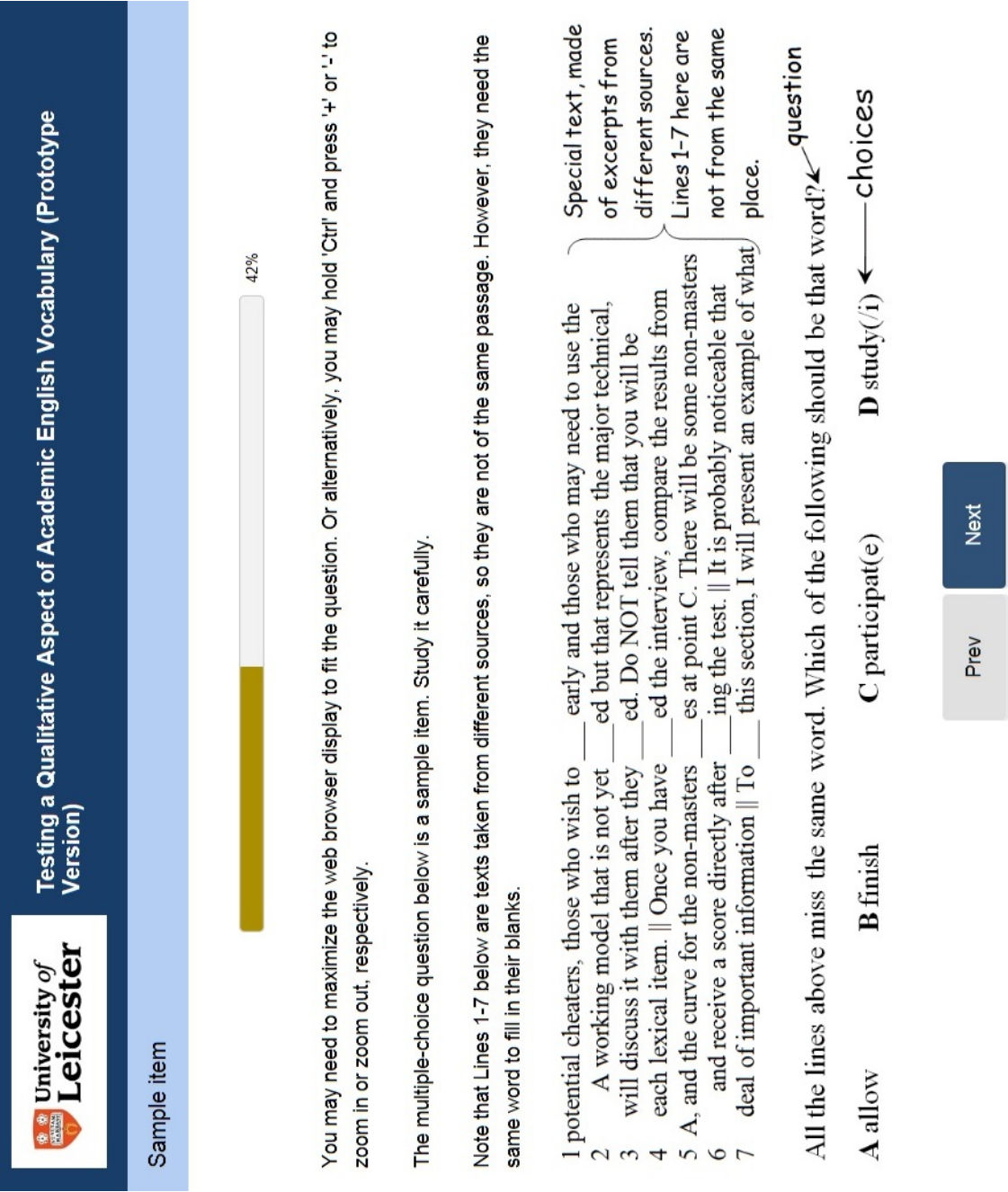


Figure 32 ConCloze 4 directions for adjusting the font size

In light of the improvements in font size, all of the respondents in ConCloze 3 and 4 seemed satisfied with the appearance of the on-screen display. Figure 31 (page 121) illustrates this, where 100% of the sample expressed positive opinions. An inference is

that the issue of font size, albeit a minor usability issue in ConCloze 2, has been adequately resolved. The size of item display in general should be appropriate for the intended population, and thus is unlikely to pose a serious construct-irrelevant threat to test presentation.

The last usability area investigated is item presentation, seeking to determine if the item design is usable for the respondents. Collected at the end of the test, the respondents' feedback is presented in Figure 33 below. Eighty percent of the responses (8 of 10) appear positive towards the one-item-per-screen presentation. Apart from Claire's and Franz's irrelevant comments, none of the respondents identified this aspect of test design as problematic. Moreover, none of them pointed out their difficulty dealing with any particular IV. These two pieces of evidence allow an inference that the item design as it stands is likely to remain appropriate for the intended population; construct-irrelevant threats from item presentation are improbable.

Test: Please give comments on test design (the overall feature, e.g., one page for one item)*.

ConCloze 2

Aaron: Clear

Björn: Good organization

Claire: Difficult

Dakota: Yes

ConCloze 3

Franz: Calculating practices may not fit to some testees

Gill: I think it was clear

Halle: I think it's suitable for this activity and it's easier to focus on each item at a time.

Igor: It looks OK.

ConCloze 4

Klavier: I think it was well structured and not too long

Lulu: Suitable

*Spelling not corrected. First letter capitalized.

Figure 33 Respondents' comments on item presentation

Thus far, this section has discussed usability issues in qualitative prototyping. Most of them are found to be unlikely threats to test-task performance and validity investigation. The only emerging concern is two ConCloze 2 respondents' inaccuracy in interpreting the concordance prompt of one item each. This has been resolved with an on-the-spot intervention and rectified subsequently with the re-introduction of a sample item. The respondents' written feedback is also collected in conjunction with observing and

assessing their responses to the test and revisions in real time. It can be contended that ConCloze 4 is likely to be usable and function well; therefore, no further iteration is needed, and the usability testing runs to ConCloze 4. The overarching claims are that (a) the construct-irrelevant threats from usability issues against the ConCloze construct representation are likely to have been minimized, and (b) the test improved could probably be satisfactorily functional with the intended population. Accordingly, the claims seem to support the hypothesis on usability (page 106): the respondents did struggle with the online testing, and the test platform is found to function well. Further, no evidence concerning usability issues can be found tied with one particular IV. An inference is that the varied features of the IVs are unlikely to cause specific usability problems.

3.3.4 Substantive content

In ConCloze 1, that the domain could be lexical-semantic is detected from patterns of responses to the options (pages 97f.). However, whether the words in the concordance lines also activate the knowledge domain remains unknown. Moreover, equally important is how the domain operates during task engagement, which is a substantive question for construct definition (cf. Table 1, page 16 for the research questions). Therefore, verbalizations of task engagement will be investigated in this section for content- and substantive-validity evidence. The focus is on whether words in the concordance lines, options, and meanings of all of the option words are likely to be processed during task engagement.

The verbal reports are elicited individually and in the same sessions as usability testing (reported on page 110). A major distinction between the two is that usability issues are mostly observed and, if necessary, dealt with on the spot, leading then to note-taking and test improvement. By contrast, verbalizations are almost always analyzed based on transcriptions. A consequence is that few salient validity features can be noted actively on-site. It would be hard, for example, to observe whether all of the concordance lines are read out and take notes of multiple usability issues at the same time. Accordingly, observing and dynamically assessing the verbalizations apply to the following two situations only. First, whenever the respondent's volume goes down or their utterance becomes mumbled, they are given a verbal nudge, e.g., 'Keep saying/talking,' 'Say loudly

please,’ ‘Keep saying whatever you are thinking/reading,’ which is intended to ensure recognizability of their verbalizations being recorded (recommended in Johnstone *et al.* 2006: para. 7; Bowles 2010: 114ff.). Secondly, when part of the content in their verbalizations appears missing but may be significant to interpreting their performance, they are interviewed semi-structuredly with immediate retrospection for clarification. The criterion for whether to interrogate retrospectively is discretion on adequacy, e.g., when an answer is picked without all the concordance lines or options verbalized, when the respondent appears confused and reluctant to choose yet is able to select the right option. In Figure 34 below, the elicitation process is visualized, where *adequacy check* represents the current dynamic assessment. Appendix 6 (page 390) offers report transcriptions.

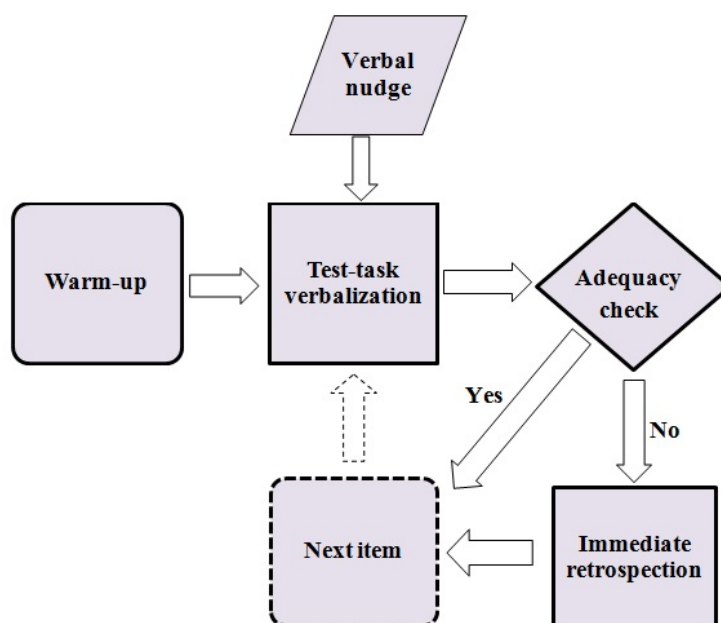


Figure 34 Session of eliciting verbal reports

Indicated in Figure 34, each entire session begins with a warm-up before a series of verbal-report elicitations. This comprises a briefing on verbalization, watching an example of how to think aloud (at, e.g., <http://www.youtube.com/watch?v=gyBYbk-gpUA>, and <https://www.youtube.com/watch?v=YDCXhDFxmM8>), and up to three think-aloud exercises (simple math). In the briefing, a point is emphasized along the lines: ‘I would like to learn *how* you get to an answer. It’s not a focus if your answer is right or wrong. Whatever you’re reading, just read it out loud. Whatever you’re thinking, just say

it out loud.” After the warm-up, the respondent verbalizes the task on each of the five IVs (cf. Table 20, page 107 for their spec). When deemed insightful, the respondent is also asked for an immediate-retrospective account thereafter (cf. Kuusela & Paul 2000 for strengths and weaknesses of types of verbalization).

It is worth stating that in actuality, it may at times be difficult to draw one-off boundaries between concurrent and immediate-retrospective accounts. This can be particularly true after a respondent comes to an unusually long pause for concentration (as in Case 1 in Figure 35 below) or seems hesitant about task engagement (Case 2). Another case is when they appear to have decided (to some extent), thereby prompting immediate interrogation on the researcher’s part, only to return to tackling the task later (Cases 1 and 3). In whichever case, when a bout of retrospection occurs during concurrent verbalization, and a response from the researcher seems required, basic protocol is followed. The protocol includes reflecting the respondent’s own thought and asking them for clarification, rather than offering clues or information potentially useful for their task completion, for instance. The aim of the reflective technique is to contain the researcher’s reactivity in test-task completion (cf. Ericsson & Simon 1993; Kuusela & Paul 2000; Bowles & Leow 2005; Bowles 2010 for challenges in analyzing verbalizations). On these accounts, the test interpretations are unlikely to be severely contaminated with construct-irrelevant variance caused by the researcher.

On Item 2 applicable

Case 1

Björn*: Hmm. [Line 6] [IA] While this solution may not be er may not be er suitable? [LP]

Researcher: Go on.

B: So, I think that answer [SP] may be choice A or B? Applicable or suitable.

R: Uh-huh.

B: Because er [LP]

R: Why?

B: But's er, but which one will be the most suitable one for the ques[tion] for all the sentences. [LP] [IA] So, erm [SP] from the last sentence I think is suitable? Answer B? [Line 7] Court held that so long as ['a law' skipped] is [E] generally [E] 'suitable'? Suitable [Line 7 cont.] and not discriminatory in being specifically aimed at religion

Case 2

James: [SP] [Line 7] Court held that no long as a law in generally blank [KWIC position] and not discriminatory in being specifically aimed at religion [LP] religion [SP] the, OK, you want me to choose the [SP] most suitable answer?

R: Yes. Now? Yeah. Keep saying whatever you are thinking.

J: Yeah, I have yeah three [SP] potential answers, and just try to submit the those key words to these blanks and see er make any sense. [Line 3] require the use of resources require that clear that [LP] so, [LP] I think er most of the cases the answer 'C' is not very suitable.

Case 3

Esther: So, same kind of question?

R: Yes, yes.

E: And the words [Options] 'applicable', 'suitable', and 'true' [SP] again, I think I look for the [IA] yeah, where the beginning is there. [Line 6] 'While this solution may not be applicable to all common property resources for those cases which may' [SP] Could be [SP] Now I see over here [SP]

R: Why so?

E: Er, [SP] 'suitable' [Option B] may not be [SP] er [SP] like suitable to use here because [Line 6] 'While this situation may not be 'suitable' [LP]

R: Uh-huh.

E: To well actually 'to all common property resources' But I think 'applicable' sounds better over here.

R: For Line 6, eh?

* E=emphasis (e.g., drawled or lengthened utterance, with spelled-out syllables), IA=inaudible, LP=long pause, SP=short pause

Figure 35 Bouts of retrospection during concurrent verbalization

In Section 3.3.3, collecting verbal reports for prototyping purposes ceases when no more usability issue is detected (cf. page 123 for the earlier discussion on this topic). The decision to stop is also supported by real-time observations of attempts to complete

the test tasks meaningfully, irrespective of differing styles of task engagement, fluency in verbalization, and (in)correct answers. Focusing on attempts to engage in the tasks meaningfully applies to all the verbalization sessions. It is observed that when concordance lines are read, option words are often also verbalized, usually in the very position of the KWIC blank. Such attempts are exemplified in Figure 35 above, where the respondents read the concordance lines and options and were able to complete the task. In fact, judged through dynamic assessment, there emerges no new finding in this respect of substantive performance as early as the beginning of ConCloze 3. Saumure & Given (2008) and Mason (2010) contended that as low as two can be sufficient as a criterion of the sample size for data saturation in qualitative studies. Using the criterion, this means that the finding on this behavioral pattern in verbalizations likely becomes saturated: whether the concordance prompt and options are read in ConCloze engagement seems to have received an affirmation. In all likelihood, these two item components are processed for task completion. The evidence also suggests that the components may contain vital clues to solving the task, so much so that all the examinees had to look for them in task engagement. Therefore, amidst variation from IV to IV, this finding can be deemed comparability in task-content processing in favor of interviant consistency. Consequently, the operational hypotheses (page 106) that these two item components are processed in ConCloze engagement are accepted.

In addition to the comparable processing of item components, interviant consistency can also be observed in item responses. Presented in Table 23 below are ConCloze 2–4 responses categorized to item number and result (correct/incorrect), in which there are the highest and lowest scorers—Lulu (five correct) and James (five incorrect)—and the rest standing in between. This pattern indicates that there is likely a competence domain acting across the IVs and accounting for this individual score variability. In other words, variation among the examinees could be interpreted as varied levels of strength in this competence among individuals. This variation could be so great, so that those with high level would likely perform well throughout the multivariant test, and poorly for the opposite (cf. pages 90f. for a similar interpretation in the quantitative analyses of ConCloze 1). Accordingly, this finding is likely to signify the potentially incremental–developmental nature of this competence and also support the inference of a core construct operating behind multiple IVs.

Table 23 Item responses and respondents

Item	Respondents with Right Answer	Respondents with Wrong Answer
1 <i>endeavor</i>	Six: Aaron, Björn, Claire, Dakota, James, Klavier	Six: Esther, Franz, Gill, Halle, Igor, Lulu
2 <i>applicable</i>	Ten: Aaron, Björn, Claire, Dakota, Esther, Franz, Gill, Halle, Klavier, Lulu	Two: Igor, James
3 <i>hypothesize</i>	Six: Björn, Claire, Dakota, Franz, Gill, Lulu	Six: Aaron, Esther, Halle, Igor, James, Klavier
4 <i>recreational</i>	Nine: Claire, Dakota, Esther, Franz, Gill, Halle, Igor, Klavier, Lulu	Three: Aaron, Björn, James
5 <i>livestock</i>	Ten: Aaron, Claire, Dakota, Esther, Franz, Gill, Halle, Igor, Klavier, Lulu	Two: Björn, James

Thus far, this section has dealt with two questions. The first is whether the concordance prompt is processed in ConCloze engagement, and the other whether the options are also processed. The other task content hypothesized to be processed is meaning of option words, which is sought in the transcriptions of the verbal reports. It is worth restating that ConCloze 2–4 collects 60 verbal reports in total (12 respondents \times 5 IVs each), ranging from approximately 3–15 minutes in duration. Considering the time constraints and the potentially large amount of data, the investigation begins with randomly selecting between Items 2 and 4, which are relatively easy among the items in Table 23 above. Easy items are focused on because if meaning of option words really counts in test-task processing, then easy items offer maximal chance of finding it verbalized by most or all the examinees. This item stratification results in choosing Item 2 (already illustrated in Figure 25, page 108). All the verbal reports for this item (12 verbal reports = 12 respondents \times 1 IV) are transcribed in this research stage. Figure 36 exemplifies some instances related to meaning of option words.

Explicit (Aaron, Dakota, Esther, Franz, Igor, James)

Franz

F: For this time, I try to *translate the choices* first, because it can be *more easier*.

Researcher: Yeah.

F: It can be *much easier*, sorry.

R: It's OK. No worries.

[...]

F: So, after *finding their meanings*, now I do take a look at certain sentence.

Igor

I: [SP] [Line 6] efficient and biological biologically appropriate while [E] this solution [E] may be may not be [SP] s[‘uitable’] OK, I decide to choose er ‘B suitable’

R: Uh-huh.

I: I think this er this question [i.e., ‘option words’] all can apply in the sentence but *the meaning is not appropriate*.

Implicit (All)

Klavier

K: Yeah, I think erm I think *that kind of fits all*. If if it fits *erm three on this*, it should be correct. I mean, [IA ‘three’] a bit left not quite enough.

R: So, why do you think the answer is not ‘suitable’ or ‘true’?

K: Erm [LP] *It kind of fits the context*, I think. Erm because it's like ‘relative stability’ and that kind of ‘a trait’ and the other [IA] share, but ‘applicable [IA ‘cause’] stability’ and being applicable *is kind of similar?* So, erm, I think that's why.

Lulu

L: Er I'll just go for the next one. [Line 2] urbanization display three research perspectives: abstract frameworks that are [E] universally [E] [SP] applicable? But at the expense of [SP] substantive specificity [LP] from this sentence I feel like it should be ‘appli applicable’. *It just er [SP] sounds right*. Haha *make sense somewhat*, you know. [LP]

Figure 36 Verbalization of meaning-compatibility check

Introduced earlier (page 125), eliciting verbalizations entails both concurrent and immediate-retrospective accounts. Figure 36 above exemplifies part of Igor's and Lulu's concurrent verbalizations, and so do all the cases in Figure 35 (page 127). The concurrent verbalizations are found in this study to often involve pauses. For example, when Lulu read up to the word *universally* in an emphatic manner, she paused verbalizing for a short while and then selected an option (*applicable*, the key). In fact, scanning through all the verbal reports seems to reveal that a frequent position for pauses is near the KWIC position (cf. Appendix 6, page 390; 60 verbal reports in total, with 12 sampled for substantive analysis in ConCloze 2–4). Because silence can be interpreted meaningfully in pedagogical contexts (King 2013), the pauses found particularly near the KWIC position are interpreted as deep processing and concentration being invoked. Such intensive processing would represent a deeper level of information processing than that being verbalized and recorded. On this account, whether the meanings of the option words

are processed for task completion may remain unclear at times in the concurrent verbalizations.

In spite of potentially deep processing in concurrent verbalizations, immediate retrospection offers a more conspicuous answer to the subtle challenge. For example, in Figure 36 above, Franz voiced his strategic plan at the very beginning of task engagement. He seemed to be saying metacognitively that learning the meanings of option words was then decisive in tackling the task. Similarly, Igor, even if getting a wrong answer eventually, explained that all the options might be used in that particular concordance line (Line 6). To him, however, the meaning of potentially some or all of the option words could be problematic. These examples of the examinees' retrospective accounts are largely about meaning-compatibility check.

In addition to explicitly referring to use of the option-word meaning, the respondents' retrospection in Figure 36 above also involves an implicit but comparable way of arriving at an answer. Implicitly processing the test tasks can be found in all the verbal reports of the sampled item. For example, Klavier tried to explain why *applicable* should be the correct answer by arguing for the compatibility of the option and a few words in context ('relative stability' and 'a trait'). Because these reports all culminate in an option being selected systematically, the evidence suggests that meaning of option words is likely to be a component of test-task content commonly used in ConCloze. Accordingly, the hypothesis that this content is processed for task completion (page 106) could also be accepted. For validity investigation, this means that lexical-semantic knowledge is likely to be a domain tested by ConCloze.

Apart from exploring the content of verbalizations, their underlying processes are also investigated. The aim is to demonstrate what the respondents do for task completion, a piece of substantive-validity evidence (cf. American Educational Research Association 1999 for processing as validity evidence). Viewing the verbalizations from another perspective, Figure 37 below depicts the relationship between the aspects of content and processes in the verbalizations: the explicit-implicit statement about meaning-compatibility check is viewed as the content interpretable in the verbalizations. Namely, the respondents went through a deep processing for test-task completion, potentially drawing on all the information they had gathered in the item. In immediate retrospection, the respondents cited agreement in meaning as a reason, for example, for choosing one

option over the others, partly in response to the task posed by the researcher (demonstrated earlier). Accordingly, the current investigation of the underlying processes is to construct processes that can account for the common behaviors across the respondents.

Proficiency in grammatical and vocabulary use

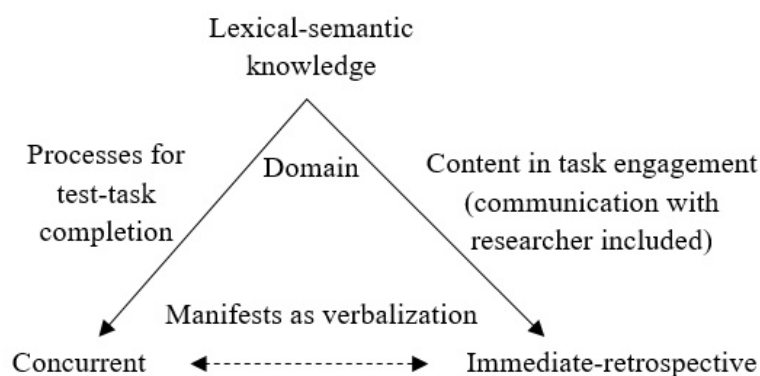


Figure 37 Verbalization as manifest ConCloze competence

Because there are no known direct studies into ConCloze processing, a Grounded-Theory approach is applied to exploring the processes underlying task verbalizations. The approach systematically generates insights out of data, rather than formulating rigid hypotheses at the beginning and testing them later (cf. Cohen *et al.* 2011). An advantage is that the data could speak for themselves, thereby maximizing the possibility of creating processes that can account for as many of the verbalizations as available. If certain processes are truly significant in task engagement, then they will withstand being constantly revised and be applicable across examinees. Figure 38 depicts this procedure, where each process label created undergoes constant checks for its capacity to explain further verbalization segments.

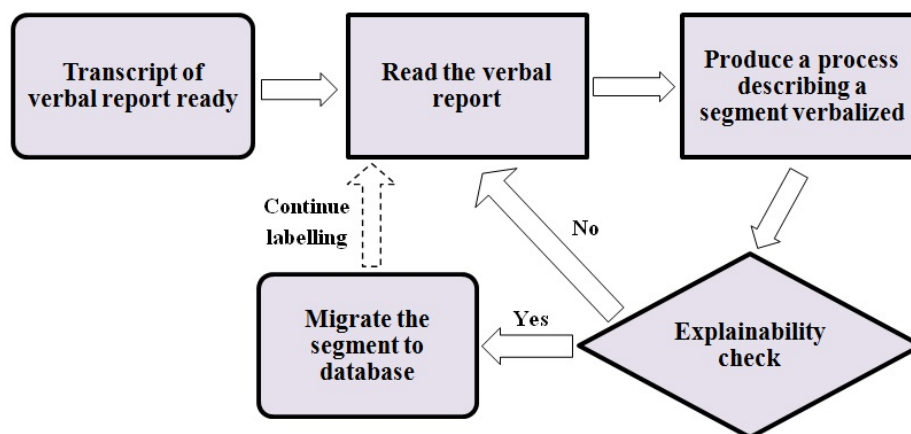


Figure 38 Tagging processes in verbal reports

In analyzing ConCloze processes, a weak form of the Grounded Theory (Armstead 2001) is used in this study: comprehensive narratives of examinees' lives are not used for constructing a theory. Rather, only a snapshot of their task processing is taken and hence the processes formulated can only represent its epistemic weak form (cf. Dillon 2013). In doing so, it uses an open tagging, which is portrayed in Figure 38 above. A tentative segment and label of a verbalization is revised over and over until best fitting newer data in hand. While the verbalization labels may not sound familiar in the general literature on language testing, they are intended to reflect the Grounded Theory-oriented nature, which could be suitable for an innovative item type like ConCloze. Each process label together with the corresponding segment are then migrated to and stored in a database, ready for aggregate processing. Illustrated in Figure 39 below, examples of the segments are arranged in order of communication units (ComUnits) as appear in the transcription.

CC	Name	Com Unit	Key	Process Level	Process/ Strategy	Verbalization
CC 2	Aaron	1	applicable	Retrospective	Assessing item components and difficulty	because until now I have answered the fourth question and according to my memory, I find that er each question is harder than the previous one
CC 2	Aaron	2	applicable	Retrospective	Assessing item components and difficulty	I don't have other clear clearer message I will just read the three choices. And then put the three words back into the sentences.
CC 2	Aaron	3	applicable	Concurrent	Focusing on clue-containing parts	[Line 1] relative stability over time and [LP] er [Line 2] are universally universally [SP] suitable eh? umm [Line 3] require you clear the [LP] er hmm [LP] applicable [SP]
CC 2	Aaron	4	applicable	Concurrent	Testing compatibility of a given word in context	[Line 1] relative stability over time and [LP]
CC 2	Aaron	5	applicable	Concurrent	Testing compatibility of a given word in context	er [Line 2] are universally universally [SP] suitable eh?
CC 2	Aaron	6	applicable	Concurrent	Testing compatibility of a given word in context	[Line 3] require you clear the [LP] er hmm [LP] applicable [SP]
CC 2	Aaron	7	applicable	Retrospective	Recognizing word associate(s)	Researcher: What word do you pay attention to? Aaron: Laws
CC 2	Aaron	8	applicable	Retrospective	Rationalizing word combinations	er in the first sentence I think the three words are all suitable for the first sentence

Figure 39 Analyzing verbal reports for processes

It is worth stating that the boundary of processes constructed may not always be clear-cut. An example is ComUnit 3 of Figure 39, in which Aaron did not read three concordance lines in full but read their shorter parts successively. This then forms a long string of prompt information which he seemed to focus on—and hence the strategy label *Focusing on clue-containing parts* for this ComUnit (to be discussed later). Because this verbalization contains phrases from different concordance lines, they are also identified

as separate ComUnits, namely 4–6. This separation is on the grounds that each concordance line could potentially be deemed an individual message where another language process can be performed. Notwithstanding these fuzzy boundaries, the tagging process culminates in a finite set of three processes and two strategies for 162 segments identified. They will be discussed in turn below together with their *post hoc* criteria. Due to the developmental nature of this study, some processes discussed may be subject to change/modification in light of new data in subsequent investigations.

Illustrated in Figure 40 below is the first process constructed: *Testing compatibility of a given word in context*. The criterion is when part or a whole of a concordance line is verbalized, usually with a sign of reactivity to the KWIC blank. Signs of reactivity include pausing near or at the KWIC blank (as in Aaron's and Esther's), and uttering the preceding word(s) in an emphatic manner (as in Björn's and Klavier's). Often, an option is also found to be inserted at the very position of the KWIC blank (as in Gill's). This process is mobilized in all the sampled verbal reports and accounts for 100% (12 of 12), taking a 52.47% majority of all the segments (85 of 162), thereby indicating its prime position in ConCloze engagement.

Name	Com Unit	Process/ Strategy	Verbalization
Aaron	6	Testing compatibility of a given word in context	[Line 3] require you clear the [LP] er hmm [LP] applicable [SP]
Björn	3	Testing compatibility of a given word in context	[Line 7 begun again] Court held [IA] so long is [E] generally [E] generally suitable? It's a [IA] applicable, I mean.
Claire	1	Testing compatibility of a given word in context	[Line 1] traits, which are characterized by relative stability over time and [SP] over time? And blank? [SP] across [E] situations? [E] positive-like capacities hmm [LP] relatively more malleable and
Dakota	7	Testing compatibility of a given word in context	[Line 3] have [E] clear [E] the suitable copy [IA 'can't be'] suitable copyright [SP] laws the suitable copyright laws? No. er
Esther	8	Testing compatibility of a given word in context	But erm [LP] OK, [Line 7] Court held that so long as a law is generally [SP] true [SP] and not discriminatory in being specifically aimed at religion [SP] er [SP] haha [laughing] I got the there's nothing more after the, so it's become difficult to understand. [SP]
Franz	14	Testing compatibility of a given word in context	Erm [LP] [Line 4] aim for a break, just for socialist reform, a reform hmm 'suitable' sounds doesn't sound good.
Gill	2	Testing compatibility of a given word in context	Let me move forward and see. [SP] [Line 2] urbanization display three research perspectives: abstract framework that [SP] are universally yeah 'applicable'.
Halle	7	Testing compatibility of a given word in context	But then again also, [Line 3 from front again] require the use of resources that require [SP] you clear the applicable [SP] copyright laws. [LP] It would be between 'applicable' and 'suitable' but for this one, I think I would go with 'suitable' maybe? [Laughing] I'm not sure. I'm gonna go to the next one.
Igor	2	Testing compatibility of a given word in context	Igor: But [Line 1] traits which are characterized by relative stability over time [SP] [IA] Researcher: Keep saying. Igor: [Line 1] and [SP] over time and [LP] Researcher: Keep saying. Igor: Applicable across situations [LP] [IA]
James	1	Testing compatibility of a given word in context	[Line 1] traits, which are characterized by the relative stability over time and [SP; KWIC blank position] across situations, positive state-like capa capacities are relatively more like malleable and
Klavier	3	Testing compatibility of a given word in context	[Line 3] require the use of resources that require you [E] clear the [E] [SP] copyright laws I think [SP] I think it's 'applicable' [LP]
Lulu	2	Testing compatibility of a given word in context	[LP] [Line 1 repeated] traits which are characterized by relative stability over time and the something across situations [LP] hmm, [LP] characterized by relative stability over time and the [LP] applicable, suitable, true [LP] hmm interesting,

Figure 40 Process 'Testing compatibility of a given word in context'

The next concurrent process is the online strategy *Focusing on clue-containing parts*. The criterion is when a concordance line is verbalized only in part, usually prior to

the KWIC position (as in Franz's verbalization in Figure 41 below). More often than not, an option is also decided upon and verbalized therein (as in Gill's and Igor's). This strategy is closely related to the first process *Testing compatibility of a given word in context* and not mutually exclusive. A distinction between them is this strategy explicitly emphasizes an element of decision-making on how to best deal with a particular situation in hand—to focus or simply read a whole concordance line. It is unknown exactly why the verbalizer decides at the moment then not to read an entire line but merely part of it. Yet, it is possible that the part focused on is meaningful for their solving the puzzle blank. This could be either because of the presence of some key words directly related to the missing KWIC in it, or because of their desire to direct concentration to the part that they believe really counts. In whichever case, this strategy seems moderate in effect size: found in 75% of the verbal reports (9 of 12), with 10.49% of all the segments (17 of 162).

Name	ComUnit	Process/ Strategy	Verbalization
Aaron	3	Focusing on clue-containing parts	[Line 1] relative stability over time and [LP] er [Line 2] are universally universally [SP] suitable eh? umm [Line 3] require you clear the [LP] er hmm [LP] applicable [SP]
Claire	13	Focusing on clue-containing parts	[SP] ‘While this solution may not [E] be [E] [SP] suitable? To all [E] common [E] property resources. [SP] or ‘this solution may not be applicable’. [SP] ‘to all common [LA ‘property resources’] for those cases which [E] may [E]
Dakota	11	Focusing on clue-containing parts	Appli applicable applicable across [SP] time [‘and ____ across situations’ skipped] positive erm sta stability over time and er ‘suitable’ across situations [SP] [Line 2] and the frameworks that are [SP] universally [SP] ‘suitable’ but at the [SP] [LA] expense of the ok? [Line 3] require the use of er resources that require you clear the [LP] er applicable applica [‘copyright’ skipped] laws before ok? [Line 4] a reform [SP] [Line 5] one three four have applicable er measurement methods sh suitable have suitable measurement methods [SP] [Line 3] have [E] clear [E] the suitable copy [LA ‘can’t be’] suitable copyright [SP] laws the suitable copyright laws? No. er [Line 6] this solution [E] may not be [E] applicable to all common [SP] [‘property’ skipped] resources [Line 7] law is [E] generally [E] [LP] [KWIC blank skipped] not discriminatory in being [LP] [Line 7, wider front part] so long as a law is generally [LP] I think A. [SP] But A and B are similar haha [laughing]
Esther	7	Focusing on clue-containing parts	So, [Line 4 partially repeated] [SP] a reform [SP] ‘true’ in the new areas [LP]
Franz	10	Focusing on clue-containing parts	[Line 3 repeated partially] that require you [E] clear [E] [LP] so clear [LA]? [SP] Right?
Gill	7	Focusing on clue-containing parts	or [Line 5 repeated partially] [LA] have suitable suitable measurement methods. [SP] 43 have [E] likely [E] [SP] eleven have potential. [LP] it turns out ‘applicable’ yeah ‘applicable’, coz ‘applicable’, it’s ‘applicable’, ‘likely’ and ‘potential’, ok. [SP]
Halle	6	Focusing on clue-containing parts	‘true copyright laws’? [SP] ‘suitable copyright laws’ [SP] might make sense. [SP]
Igor	11	Focusing on clue-containing parts	[The chunk rephrased] suitable in the new a[‘rea’] This one should be ‘suitable’ eh [laughing] [LP]
James	7	Focusing on clue-containing parts	[Line 6 restated] may not be blank [KWIC position] all common property resources, for those of cases may [SP]

Figure 41 Strategy ‘Focusing on clue-containing parts’

With regard to immediate retrospection, the first process constructed is *Rationalizing word combinations*. The criterion is when the respondent tries to justify their decision made to the researcher, either on their own or upon interview (as in Esther's verbalization in Figure 42 below). This usually entails explaining why one option should be the answer, e.g., by means of rejecting another option (as in Aaron's), clarifying their word of choice (as in Dakota's), describing context of use for the choice (as in Franz's). Typically, this process exhibits reactivity the words in a concordance line have towards the KWIC, and hence *word combinations* in the designation. For example, Claire pointed out that the distractor *true* does not go well with the phrase *copyright laws* in Line 3 (see Figure 25, page 108 for the item). Another example is when Gill pointed out that the distractor *suitable* does not go well with the adverb *generally* in Line 7. Nonetheless, despite their attempt to justify their answer, it can sometimes be a challenge for the respondents to articulate why one option would be more appropriate than the others. For example, in Figure 42, Claire, Dakota, Igor, and Lulu coincided on using the somewhat ambiguous expression *make sense* in order to rationalize their decision. Yet, they did not specify the language domain involved in distinguishing one option from the others. This process can be found in 92% of the sampled verbal reports (11 of 12), accounting for 19.14% of all the ComUnits identified (31 of 162), a substantial proportion among them.

Name	Com Unit	Process/ Strategy	Verbalization
Aaron	9	Rationalizing word combinations	when I read the three subquestion [i.e., Line 3] I think 'laws' are already real ones and I'm sure to this type laws I think 'true' repeatable for me. Maybe it's not appropriate.
Claire	4	Rationalizing word combinations	Researcher: Uhhuh, ok. Why? Claire: I think er the copyright laws? [SP] Yeah, because you don't [IA] now the copy now near [SP] 'the copyright laws' Researcher: Uhhuh. Claire: So [SP] I think er if I I put [SP] 'true copyright laws' I think er it doesn't [IA] make sense.
Dakota	14	Rationalizing word combinations	Researcher: Any key word? Any words that you think are Dakota: Yeah, I just the I just the put the words here if the if that makes sense to the Researcher: What do you mean, 'make sense'? Dakota: 'Make sense' is is means like er you read like er sentence like er they did er they [IA 'mean er'] something. They did the say something. But actually these two words [Option A 'applicable and Option B 'suitable'] two words quite close I mean. [LP]
Esther	13	Rationalizing word combinations	Esther: but over here, it's also like because er [SP] 'true', 'suitable', and 'applicable', these words are kind of, you know, having more or less the same meaning or like closely related, you know. Researcher: Yeah. Esther: Related; they are not opposite to each other.
Franz	7	Rationalizing word combinations	Because when you do research erm it tends to er [SP] to the situations that you must apply something [SP]
Gill	13	Rationalizing word combinations	If you are a student of English, you should know that 'generally' it just be can't 'suitable', but it is best to be 'generally applicable'. Researcher: So, you think Line 7 helps you the most to clear-cutly decide. Gill: Yeah.
Halle	10	Rationalizing word combinations	because if it would have been 'suitable', it would have been 'suitable [E] for [E] something', so I'm gonna go with 'applicable'.
Igor	5	Rationalizing word combinations	I think the last one [i.e., option; Option C 'true'] is not make sense.
James	20	Rationalizing word combinations	but the but the' [SP] I think it's er if I er 'applicable' and 'suitable' are equally ok, but er according to my experience, the 'suitable' is er most language some of this word. That's why I use the. Compared to this 'true', these two, giving good sense.
Klavier	8	Rationalizing word combinations	Researcher: What about the other lines? Are there any key words that help you specially? Klavier: erm [LP] this one, [Line 4] 'a reform applicable in the new areas' erm [SP] 'suitable' wouldn't be correct. I don't think it's true as well. Just wouldn't be that's erm because you can apply them in the new areas. erm
Lulu	5	Rationalizing word combinations	[LP] from this sentence I feel like it should be 'appli applicable'. It just er [SP] sounds right. haha make sense somewhat you know. [LP] If you talk about the [SP] framework, you talk er it is also it is almost like talk about framework. The connotation of 'framework' is almost, you know, you apply framework. [SP] Right? So you say whether framework is applicable. You don't say whether er a framework is true. Erm you may say whether framework is suitable. But it's suitable to some kind of situation, it's almost like you need a erm [SP] you need a er another erm phrase a, an, the for the qualify 'suitable'. So, it feels like it should be 'applicable'.

Figure 42 Process 'Rationalizing word combinations'

The next immediate-retrospective process is *Recognizing word associate(s)*. The criterion is when a respondent picks individual words or short phrases from the concordance lines, mostly in order to support their decision or answer. For example, in Figure 43 below, Aaron reported that seeing the phrase *measurement methods* (in Line 5; see Figure 25, page 108 for the item) made him think of the option word *applicable*, which is also the key. Similarly, Franz saw the word *research* in Line 2 and thought of *applicable*. In less conspicuous cases, when interrogated for the clues they use to reach a decision, the respondents may take a short phrase, usually encompassing the KWIC position, in the way as if they are aware that the phrase may contain important clues. For example, in Figure 43, when Igor was asked about words that had helped him to arrive at an answer, he vaguely re-verbalized a predicate part of Line 6 with the chosen KWIC filled out. All this evidence seems to underpin an earlier inference (page 128) that the concordance prompt contains important clues to solving ConCloze tasks. Irrespective of whether word associations are verbalized conspicuously or subtly as part of a concordance line selectively verbalized, the process is found in 75% of the sampled verbal reports (9 of 12), garnering 11.11% of all the ComUnits identified (18 of 162).

Name	Com Unit	Process/ Strategy	Verbalization
Aaron	10	Recognizing word associate(s)	And then I look at the fifth question has 'measurement methods' and in my memory I think methods measurement this kind of word I usually be described by 'applicable'. Also, my first impression is 'applicable'.
Dakota	13	Recognizing word associate(s)	Researcher: Aha. Anything else that helps you to get to the answer? Dakota: [Line 2] are [E] universally [E]
Esther	9	Recognizing word associate(s)	[Line 7] Law is 'applicable' even erm 'applicable' can be used with 'the law',
Franz	6	Recognizing word associate(s)	Ah, interesting. [SP] When I see the word 'research' [SP] my mind links to the word 'applicable' [Option A] [SP]
Gill	4	Recognizing word associate(s)	Yeah, because it it requires you to do something in order to [SP] because of the words 'copyright laws', I guess. So, applicable.
Igor	17	Recognizing word associate(s)	Researcher: What what do you mean? It helps you with [SP] What are the key words in Line 6 that helps you? Igor: [Line 6] 'may not be suitable to all common property' [the KWIC added] This one, yeah, it's look [SP] common. It's commonly used. I don't know I just guess, yeah.
James	16	Recognizing word associate(s)	[SP] [Line 6] 'applicable measurement ['methods' skipped],
Klavier	5	Recognizing word associate(s)	Because erm it can be like 'applicable copyright laws' and erm it can be 'applicable across situations'
Lulu	6	Recognizing word associate(s)	[LP] from this sentence I feel like it should be 'appli applicable'. It just er [SP] sounds right. haha make sense somewhat you know. [LP] If you talk about the [SP] framework, you talk er it is also it is almost like talk about framework. The connotation of 'framework' is almost, you know, you apply framework. [SP] Right? So you say whether framework is applicable. You don't say whether er a framework is true. Erm you may say whether framework is suitable. But it's suitable to some kind of situation, it's almost like you need a erm [SP] you need a er another erm phrase a, an, the for the qualify 'suitable'. So, it feels like it should be 'applicable'.

Figure 43 Process 'Recognizing word associate(s)'

The last strategy constructed is *Assessing item components and difficulty*. The criterion is when the verbalization reflects their general strategy in dealing with the task and item components and is usually involved with meta-cognitive evaluation. For example, Figure 44 below shows Aaron's and Franz's verbalizations highlighting that the options were to be read prior to any concordance line. On other occasions, the reflection is more subtle, though. For example, Esther must have evaluated the overall item format quickly, and was only heard asking if the question was of the same type as the previous item (Item 1). Similarly, Halle did not verbalize how she was going to tackle the task altogether but simply started reading the whole first concordance line and all the options thereafter. In other words, she attempted to fill out the first concordance line before continuing to the rest, rather than reading all the lines and then the options in order of

presentation. In this way, she very likely had improvised her own style of ConCloze engagement online.

Irrespective of whether the strategic engagement of a task is performed explicitly or subtly as inferred from a variation in how the task is engaged with, the strategy *Assessing item components and difficulty* can be found in 67% of the sampled verbal reports (8 of 12). Nonetheless, the occurrences of this strategy account for only 6.79% of all the ComUnits identified (11 of 162), constituting a relatively small proportion to all the processes constructed. It can thus be regarded as not greatly significant to ConCloze processing. Moreover, Alderson (1990) investigated verbal reports of a reading comprehension test and found that individual examinees may vary in how they mobilize their strategies in approaching different items. This implies that the small proportion of this strategy may be due to it being generic to tests with a reading element. The strategy may thus be argued as peripheral in terms of construct relevance, and could be deemed part of test-format and research-method variance and thus only an artifact of discursal construction.

Name	Com Unit	Process/ Strategy	Verbalization
Aaron	2	Assessing item components and difficulty	I don't have other clear clearer message I will just read the three choices. And then put the three words back into the sentences.
Esther	1	Assessing item components and difficulty	Esther: So, same kind of question? Researcher: Yes, yes. Esther: And the words [Options] 'applicable', 'suitable', and 'true' [SP] again, I think I look for the [IA] yeah, where the beginning is there.
Franz	1	Assessing item components and difficulty	For this time, I try to translate the choices first, because it can be more easier... Erm, well, [Instructions] 'All the lines above miss the same word. Which of the following should be that word?' Erm, applicable, suitable, and true. They are all different. [SP] Right?...So, after finding their meanings, now I do take a look at certain sentence.
Halle	1	Assessing item components and difficulty	[Line 1] The first line is 'traits, which are characterized by relative stability over time and blank' across situations, positive state-like capacities are relatively more malleable and [SP] and the words I have are applicable, suitable, or true?
Igor	1	Assessing item components and difficulty	You have to cut something that is 100% wrong out, and ok your chance to get the right answer that higher. From four choice you get three. That's mean the percent to correct is increase.... Ah, this one is [SP] has three, ok. [Option A] applicable, [Option B] suitable, [Option C] true. OK, this meaning is no problem.
James	9	Assessing item components and difficulty	Yeah, I have yeah three [SP] potential answers, and just try to submit the those key words to these blanks and see er make any sense.
Klavier	6	Assessing item components and difficulty	Researcher: So, you have read just three lines and then you got the answer? Klavier: Yeah, I think erm I think that kind of fits all. If it fits erm three on this, it should be correct.
Lulu	3	Assessing item components and difficulty	hmm interesting, because I was expecting maybe a noun here in this blank. But somehow all these are adjectives. [SP] er make me feel the they don't feel like they are parallel structures in these sentences. hmm OK. [SP] er I'll just go for the next one.

Figure 44 Strategy 'Assessing item components and difficulty'

In summary, several questions regarding task content and substantive performance are answered in this section. The first issue is whether the concordance prompt and item options are really processed for task completion. This has been addressed by the examinees' comparable verbalization of the item components. The second question is whether the meaning aspect of the options is also processed for task completion, which has been answered by the examinees' referring to meaning of the options in their verbal reports. The third question is whether there are core processes that most or all of the examinees mobilize in ConCloze. This has been dealt with by constructing three processes and two strategies in a Grounded Theory-oriented fashion, the proportion of which is summarized in Figure 45 below. All this initial evidence distinguishes substantive similarity from varied idiosyncratic styles, leading to a sharper definition of

the construct. An argument in this research stage is that the construct competence may include lexical-semantic knowledge activated by clues in the concordance prompt and mobilized through testing their compatibility in meaning with those of the available options.

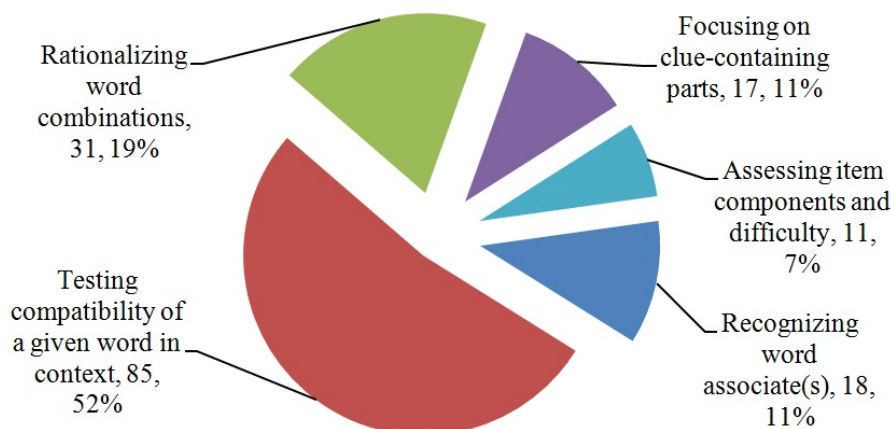


Figure 45 Test-taking processes and strategies in ConCloze 2–4 Item 2

3.3.5 Conclusion and decision

In this part, there have been several inferences made based on empirical evidence from responses to ConCloze 2–4. They are summarized in Table 24 in no particular order.

Table 24 Major evidence and inferences

Entry	Evidence	Inference
1	Test improved through rapid iterations	Usability issues minimized and ConCloze performance reflecting the construct more accurately
2	Concordance prompt, options, and their meanings verbalized by respondents	Construct domain tapped into by these elements of test-task content
3	Respondents' performance varied by items	Construct of incremental–developmental nature
4	Common processes constructible out of verbal reports	Substantive evidence of construct validity with increased effect size
5	No usability issues attached to one particular IV	Intervariant consistency across IVs, and hence core construct shared by IVs

ConCloze 2–4 are each composed of five items generated based on multiple IVs. A primary finding is that no usability issue is found to be connected with any of them.

For example, a three-option IV (Item 2; cf. the spec, page 107) and four-option IVs (the remaining items) do not cause specific usability problems. This finding has two implications for this study. First, core processes underlying variable content of the IVs may exist, such that task engagements would be comparable across these IVs. Secondly, the test platform does not cause problems of test presentation and task specification, meaning that response invalidity caused by usability issues has been mitigated and the responses can represent the construct-relevant performance.

In tandem with the usability testing, substantive processing is also investigated through the test of multiple IVs. Because each concordance line is unique, and so is the concordance prompt, each item may be considered a unique stimulus given to the examinees. As such, another key finding in this part is that the item components processed in task engagement are found to be comparable in spite of the IVs, and so are the underlying processes. Therefore, the comparability may be deemed consistency in substantive processing across different stimulations (cf. Schneiderman 1980 for significance of behavioral consistency amidst situational variability).

Given the response validity and substantive consistency, the hypothetical construct could be appraised meaningfully. In Chapter 1, the test purpose has been set hypothetically to serve as a proficiency test on professional and academic English grammatical and vocabulary use (page 22). Then in Chapter 2, a lexical network in the item prompt has been hypothesized to be formed by the associations of the target word in the task (page 50). In ConCloze 1, initial evidence seems to indicate that proficiency in vocabulary use, particularly in relation to lexical meaning of option words, may be involved in task processing. In the current ConCloze 2–4, the words in the concordance prompt are also found to likely be processed for task completion. While it is still unclear if the words in the prompt form a lexical network for the target word, they are likely to be used for assessing the likelihood of compatibility with one particular option against the others. Therefore, it may be claimed that another construct process of the proficiency in vocabulary use is to choose words that are appropriate for the contexts.

Notwithstanding being able to refine the dimension of vocabulary use, no distinguishable item responses can yet be observed so as to refine the dimension of grammatical use in the test purpose. In sum, initial findings suggest that the ConCloze

item type may test proficiency in vocabulary use, especially by choosing words based on their meanings to fit contexts properly.

Based on the findings in this version, substantive-validity evidence might be argued as likely to be sufficient for the qualitative prototyping. When considered in tandem with the quantitative evidence from Part 3.2, it makes triangulated efforts in arguing that the ConCloze item type can likely be used for measuring a discrete competence, which currently is found to involve lexical-semantic knowledge invoked by testing compatibility of a given word in the concordance context. Given this likelihood, that ConCloze elicits responses reflective of no underlying competence could also be ruled out. For more construct-related inferences, it is decided that this validity investigation should proceed into larger-scale testing.

3.4 ConCloze 5: Field-testing

3.4.1 Rationale

In the previous part, the prototyping phase of this study is completed. The investigation continues in this part by field-testing ConCloze on a larger sample of examinees, which can be useful for three reasons. First, in statistical terms, the samples in the quantitative–qualitative prototyping are relatively small. While they are sufficient for prototyping purposes, only a limited number of analyses can be performed thereupon. For example, a key quantitative measure in ConCloze 1 is Cronbach’s alpha coefficient, which essentially indicates the ratio of inter-item covariance to total-score variance. Its drawback is that a high coefficient can result from a high total-score variance, which can occur without the scale being unidimensional. While alphas-if-item-deleted are also investigated to get around this mathematical flaw, its small sample size risks an underpowering effect in statistics—so small that significant variances may not be distinguishable. Thus, in this stage of research, a larger-scale test administration allows wider categories of analysis, particularly inferential ones. A bigger body of evidence can accordingly be gathered before the main stage of test use (ConCloze 6), thereby increasing confidence in drawing construct-related inferences for the validity argument.

Secondly, in nomological terms, the prototyping phase establishes that lexical-semantic knowledge is a likely part of the construct proficiency for the ConCloze item

type. It also finds that the construct domain is potentially incremental–developmental in nature. Yet, it cannot determine whether demographic factors, e.g., age, L1 background influence score variability significantly. A scenario could be that, for example, age is a primary contributor to ConCloze performance, such that lexical-semantic knowledge is only its corollary. Therefore, collecting a wider range of responses can ward off uncertainty as to whether there exists such association. It can determine if demographic variables are construct-relevant, merely confounding variables, or entirely construct-irrelevant. This can then result in a deeper understanding and thus sharper definition of the construct.

Lastly, in heuristic terms, the corpus-based frequency level of the target words is ruled out in ConCloze 1 from being the sole difficulty driver (e.g., page 96). However, relationships among item components and different elements of test-task content remain largely unknown. A possibility could be that, for example, the number of options marginally counts for determining item difficulty whereas semantic relationship among them is comparatively more important. Therefore, analyzing quantitative item responses from multiple IVs (to be discussed later) can fine-tune the previous finding by placing these IVs in a continuum of difficulty effects, if possible. This in turn can indicate what element in each test task would likely count the most in ConCloze processing. For example, number of options may be found the least effective difficulty driver, when compared with the other elements of test-task content, which would then indicate its small effect on substantive processing. In sum, the difficulty drivers identified will serve as surrogates for the linguistic areas processed in the tasks.

For the reasons outlined, the item type will be explored quantitatively in this part by first adjusting the old test spec from ConCloze 1 in Section 3.4.2. Then response analyses follow in Section 3.4.3, where demographic factors are hypothesized not to have significant effects over ConCloze-score variability. Another hypothesis in Section 3.4.3 is that different elements of task content affect item difficulty differently. Specifically, lexical-semantics of the option words is found in ConCloze 2–4 to likely be processed by all the respondents (cf. page 131). Therefore, if those meanings of the option words are made such that they are easier for engaging in, e.g., by means of obvious differences in meaning (i.e., semantic distance), then it may be found that the test tasks would be easier as the examinees can easily distinguish the target word from the distractors. This means

that the latter hypothesis from above can be operationalized as: the IV whose semantic relationship among the options is distant or unrelated will be the easiest one.

In addition to distancing the options semantically, another element of task content is also worth experimenting for substantive evidence. ConCloze 2–4 identify the major underlying process as *Testing compatibility of a given word in context* (page 135), meaning that the amount of context clues can potentially be decisive for testing the compatibility. Therefore, another interim hypothesis is an IV with five words on either side of the KWIC is the most difficult one as this IV, when compared with the others, is likely to carry the least word information functioning as context clues to the KWIC. Moreover, for construct definition, other construct-related issues will also be addressed in Section 3.4.3. Finally, this part ends in Section 3.4.4 with making an informed decision on test length and in Section 3.4.5 with a summary of construct inferences.

3.4.2 Test spec

In ConCloze 2–4, multiple IVs have been trialed through rapid iterations of usability testing and verbalization analysis. An idea in this part is to transform ConCloze 1 items using the features of those IVs trialed. Transforming old items can be useful for three reasons. First, as introduced in Section 3.4.1 (pages 147f.), differing effects of these IV features on item difficulty, if any, will likely become measurable through the present larger sample size. This can be significant to the validity argument, as design features that are able to determine or predict difficulty can be used as construct-validity evidence (Haladyna 2004; Hoffman *et al.* 2006; Mislevy 2007). Modifications in item features that lead to varied ConCloze performance mean that such features have a role to play in task processing and hence the construct.

A second reason in support of the item transformation is randomization of difficulty through multiple IVs. In theory, multiple IVs may involve more varied substantive content than a single IV. When the overall variance of a scale is high, the scale facility will likely get moderated (i.e., the items become generally moderately difficult). Accordingly, varying item features can be a way to improve ConCloze in terms of average difficulty. It is worth noting that the ConCloze 1 scale, made of univariant items, has been evaluated to be inappropriately difficult (pages 102f.). With multiple IVs

introduced here, it is likely that the test difficulty will be lowered and better match the average level of proficiency of the target examinees.

Lastly, developing a test can be a costly and time-consuming enterprise. Considering this, modifying and reusing the items generated in the prototyping can be resource-wise. In doing so, it also indicates malleability of the ConCloze item type to suit particular needs, meaning that the present item transformation could convey a message for test use. As introduced in the rationale (pages 1f.), the message is that ConCloze could have potential for item generation in the testing industry.

Given that the direction of item design has been outlined, the test design of field-test items can be described in Table 25 below. Entry 1 presents criteria based on previous item statistics in Table 16 (page 94) which are used for selecting and turning items from ConCloze 1 into items featuring particular IVs. For example, in Figure 46 below, Item 19 in this version is contrasted to the corresponding precursor Item 24 from ConCloze 1. Item 19 represents IV4, featuring (a) ten concordance lines, (b) ten words on either side of the KWIC position, and (c) three semantically related options, with at least two synonyms. Appendix 3 (page 334) provides the items of the current test version.

Table 25 Guiding language for ConCloze 5

Entry	Guiding Language (Test Design)		
1	The test has 30 items, the features of which can be described below.		
	ConCloze 1 Item*	Previous Statistics	ConCloze 5
	5, 7, 11, 19, 35	IF 0.3–0.39, ID \geq 0.2	IV1: 7 lines, 10 words, 4 unrelated/distant options
	3, 10, 27, 36, 39	IF \geq 0.7, irrespective of ID	IV2: 7 lines, 5 words, 4 related options
	1, 4, 15, 33, 34	IF 0.4–0.59, ID \geq 0.2	IV3: 7 lines, 10 words, 3 related options
	6, 13, 18, 24, 26	IF 0.3–0.69, ID $<$ 0.2	IV4: 10 lines, 10 words, 3 related options with at least one synonym
	23, 25, 29, 32, 38	IF 0.3–0.39, ID \geq 0.2	IV5: 10 lines, 10 words, 4 related options
	12, 14, 21, 28, 37	IF 0.6–0.69, ID \geq 0.2	IV6: 7 lines, 10 words, 4 related options
2	The six IVs are arranged in the test in numerical order, with five items each. For example, IV4 is represented by Items 16–20. Another example is in IV6, where Item 12 from ConCloze 1 becomes Item 26 in ConCloze 5.		

* For example, Items 5, 7, 11, 19, and 35 in ConCloze 1 become Items 1–5 (IV1) respectively here. They feature seven concordance lines, ten words on either side of the KWIC position, and four semantically unrelated or distantly related options (the key included).

Item 24

ConCloze 1



1 the marketplace. We are often asked to rejoice that the _____ 4% success rate is so high. If low success is
 2 culture to culture, however, subject to concepts of individualism, the _____ aesthetic, and a host of social or ritual beliefs particular
 3 way to a salon. There are many methods, some even _____ almost since the birth of photography. The first method is
 4 instruction. We failed to appreciate how deeply entrenched were the _____ attitudes about what constituted appropriate social studies instruction for the
 5 quarrel about their theoretical compatibility, these were two sets of _____ ideas that were very powerful in American management. The first
 6 focusing entirely on the dining-room dimension of language. Thus the _____ linguistic point of view ends up indistinguishable from that of
 7 you think you have made an important discovery counter to _____ opinion? Over lunch a number of years ago, one senior

All the lines above miss the same word. Which of the following should be that word?

- ☐ A current
- ☐ B indicative
- ☐ C prevailing
- ☐ D standardized

Question 19

ConCloze 5



1 the marketplace. We are often asked to rejoice that the _____ 4% success rate is so high. If low success is
 2 culture to culture, however, subject to concepts of individualism, the _____ aesthetic, and a host of social or ritual beliefs particular
 3 way to a salon. There are many methods, some even _____ almost since the birth of photography. The first method is
 4 instruction. We failed to appreciate how deeply entrenched were the _____ attitudes about what constituted appropriate social studies instruction for the
 5 need to be put off by e-discovery. In fact, the _____ court rulings have created standards close to the Federal Rules
 6 quarrel about their theoretical compatibility, these were two sets of _____ ideas that were very powerful in American management. The first
 7 focusing entirely on the dining-room dimension of language. Thus the _____ linguistic point of view ends up indistinguishable from that of
 8 contemporaneity, leading one paradigm rather than another to become the _____ model. Once a new paradigm takes hold there is an
 9 you think you have made an important discovery counter to _____ opinion? Over lunch a number of years ago, one senior
 10 already contained in existing human rights instruments. Capitalizing on the _____ sentiment at the Copenhagen meeting that the time was ripe

All the lines above miss the same word. Which of the following should be that word?

- ☐ A current
- ☐ B indicative
- ☐ C prevailing

Figure 46 Transforming a ConCloze 1 item to a ConCloze 5 one

The criteria outlined in Table 25 above are intended to improve the old items. On the one hand, ConCloze 1 has a small sample size and its item statistics can only be deemed tentative. Yet, those statistics from ConCloze 1 could give an approximate idea about the quality of the items, which may suggest a suitable modification for item improvement. For example, the criterion for selecting a precursor item for IV4 is one with an IF between 0.3 and 0.69, and an ID lower than 0.2 (cf. Table 16, page 94 for ConCloze 1 item statistics). This means that each precursor item like Item 24 in Figure 46 above has

had a suitable difficulty level but poor discriminability in the prototyping. An implication is that some low performers happen to score, probably by chance or partial knowledge whereas some high performers fail to score. It thus follows that a modification that can give more context clues and a higher success rate to those high performers may improve the item. In Figure 46, three concordance lines are therefore added to the original prompt of Item 24, intended to give the examinees additional clues. One old distractor ‘*standardized*’ is also arbitrarily discarded, intended to give them a higher success rate. As the distractor *current* is already in a paradigmatic relationship (near-synonymous) with the target word *prevailing*, it is retained as now appears in Item 19 of this version. For convenience, effects of item modifications hypothesized are summarized in Table 26 below, where corresponding values from ConCloze 1 function as bases for assigning a modification in this version.

Table 26 Expected effects of modification on ConCloze 5

ConCloze 1		ConCloze 5	
Evaluation Base (\bar{x})	Primary Modification	Expected Result	Observation Area
Cronbach's α 0.84	Test shortening	Still acceptable, ≥ 0.7	Test level
Average IF (0.45)	Multivariant	Middling, > 0.45	Test level
IF 0.3–0.39 (0.38)	IV1: unrelated/distantly related options	Higher IF	Items 1–5
IF ≥ 0.7 (0.78)	IV2: 5 words on either side	Lower IF	Items 6–10
IF 0.4–0.59 (0.5)	IV3: 3 options	Higher IF	Items 11–15
IF 0.3–0.69 (0.52), ID < 0.2 (–0.03)	IV4: 10 lines, 3 options	Higher IF, satisfactory ID	Items 16–20
IF 0.3–0.39 (0.31)	IV5: 10 lines	Higher IF	Items 21–25
IF 0.6–0.69 (0.63)	IV6: N/A	Retained	Items 26–30

Another example of item transformation is IV1 in Table 26 above. The criterion looks for a ConCloze 1 item with an IF between 0.3 and 0.39, and an ID over 0.2. Those item qualities indicate that each original item from ConCloze 1 already possesses satisfactory discriminability but is somewhat difficult. In Section 3.3.4 (page 131), option-word meaning is found to be processed for task engagement, suggesting that semantic relationship among them could be an item-difficulty driver. Because the quantitative prototyping uses semantically related options (cf. page 71 for their spec), it thus follows that if their semantic components become unrelated or only distantly related, they should be easier to distinguish. To lower item-difficulty level, IV1 is hence designed to have semantically distant options.

The quantitative prototyping does not elicit responses to items with semantically related options (page 71). Then the qualitative prototyping obtains only a limited amount of responses for an IV with semantically distant options (i.e., its Item 4; cf. page 107 for the spec, and page 129 for the quantitative analysis). This means that thus far, there has been no sizable amount of item responses ever obtained in relation to semantically distant options. Designing and field-testing IVs like IV1 discussed above is therefore important because it could address the inadequacy by providing quantifiable evidence regarding semantically distant options for the first time in this research.

In addition to informed item transformation, the spec also considers appropriateness of test length. ConCloze 1 has 39 items, which has been evaluated as inappropriately long (cf. pages 86 and 100, for instance). Even though a long test usually enjoys higher test reliability than a short one, an implication for the current ConCloze version would be that there is room for it to improve in this respect. Lowering test length is intended for more pleasant test-taking experience and hence less response invalidity, if any. The spec in Table 25 (page 150) deals with this in its Entry 1, which arbitrarily lowers the total number of test items to 30.

In addition to the construct-related issues, two peripheral issues also need tackling via test design. In ConCloze 1 (page 89) and ConCloze 2–4 (page 112), data on previous standardized English test scores are collected as part of the demographic descriptions of the respondents. The first issue is that many of these respondents had not taken any of such tests in recent years. For those who had, their results spanned across a long period of time, namely, over a decade in ConCloze 1 and approximately 15 years in ConCloze 2–4. Consequently, the data in this respect are limited in effect size and difficult to compare validly across individuals. Moreover, up to this stage of research, part of the ConCloze construct is found to be at least lexical-semantic knowledge, a domain in human-language faculties known to be dynamic (e.g., Schmitt 2010: 155). Even in a short time span of a few years, the level of the knowledge can change, suggesting that the standardized-test scores reported may not reflect the respondents' current state of the construct domain.

Considering the non-reflective linguistic backgrounds collected, self-ratings are introduced in this version as part of the demographic questions. Seeking to obtain self-ratings of English skills can be useful because it gives identical data on the respondents

that could be related to their English proficiency. Illustrated in Figure 47 below, the self-rating questions are placed prior to the old question about scores from standardized English proficiency tests. On the one hand, it is acknowledged that the self-ratings may not allow a reliable comparison of actual language proficiency across individuals. Yet, they do provide the most up-to-date and readily comprehensible evaluation of the respondents' own English mastery level. In Figure 47, the evaluation is enumerated along a scale of very good–poor across the four discrete skills. This means that the self-ratings provide identical, up-to-date and simple information relating to the English proficiency that can be collected simultaneously together with ConCloze responses.

Self-ratings are empirically found to have face validity and high internal consistency (Bachman & Palmer 1989). Accordingly, in addition to the usefulness highlighted above, the respondents' self-ratings are also likely to validly reflect idiosyncratic perceptions of their English skills (cf. Luoma & Tarnanen 2003 for concepts underlying self-ratings). An argument could be that self-ratings, albeit not always comparable across respondents on their very proficiency in general English, are individual and personal in nature and therefore can be used as an identical piece of demographic information. Analyzing this information can essentially help to ward off uncertainty as to the association of ConCloze scores, or a lack thereof, with demographic variables. This means that the self-ratings are treated in this analysis not as a linguistic variable but as a non-linguistic variable that can assist with defining the linguistic nature of the item type.

How would you rate your own English ability?

	Very Good	Good	Fair	Poor
Speaking	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Listening	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

When did you last take a standardized test (such as TOEFL, IELTS, TOEIC, CAE)? (State the month and year, e.g., Sept 2012. If no such previous test at all, write N/A.)

What is your score in that test? (If none, write N/A.)

Which? (CAE, FCE,
KET, PET, CPE,
IELTS, TOEFL iBT,
TOEFL CBT, TOEIC
etc.)

Score?

Figure 47 Collecting self-ratings of English proficiency

Peripheral and unforeseen, the other issue needing tackling via test design is device for test delivery. It is recognized in this stage of research that a personal computer—laptop or desktop—may not be the only category of device through which respondents take the online test. Rather, a greater possibility that comes with a larger sample is that some respondents may use mobile devices, e.g., smartphones, tablets. Ideally, a solution is to make the display of test questions functional for all such devices, be it auto-rotated horizontally or vertically. However, complex technical difficulties could yet arise in test construction and presentation. For example, the very nature of a concordance prompt is that when displayed vertically on a smartphone, the concordance words may not be legible without zooming in. When the prompt is zoomed in, co-text words far from the KWIC position would simply fall out of sight, which means the left-hand and right-hand parts of the concordance could not be viewed simultaneously. Moreover, if a function of left–right–up–down swipe is encouraged in test-task engagement, the respondents may suffer severe fatigue from straining their eyes to read multiple tasks consecutively in such a small space. All these examples imply that there

are potentially a large number of issues concerning designing ConCloze to fit multiple types of testing device. This could be especially true pertaining to many different makes and configurations of mobile devices, so much so that technical complications to accommodate great variation in smart devices may overshadow the present focus on validity investigation.

In light of the possibility of use of mobile devices, a recommendation is added to the test-introduction page, saying that the test is not designed for mobile devices such as smartphones. On the one hand, it is acknowledged that this may cause some inconvenience to those respondents who find and click the test link on their mobile devices in the first place. Yet, given that the test is open reasonably long (approximately one month, available at <https://www.surveymonkey.com/s/AcadVocabTest>), this recommendation should help with logistic issues, so that the test platform can provide the best possible testing experience with the test resources available. The recommendation is illustrated in Figure 48.

University of Leicester

Testing a Qualitative Aspect of Academic English Vocabulary (Field-test Version)

Exit this test

Hi, there!

3%

Want to be part of a new vocabulary test format being developed? Would you like to have the quality of your own academic English vocabulary tested? If your answers are all 'Yes', this is a place for you!

My name is Kunlaphak Kongsuwannakul, a doctoral student at the University of Leicester School of Education. Thank you very much for joining my research. It is about developing new types of question for testing academic English vocabulary. Your responses will be really helpful.

In this test, you will be asked first to give some basic information about yourself for score-reporting purposes. Then you will find one sample question; study it carefully. Thirty multiple-choice questions will then follow. At the end, you will give your email address, to which you will get your score reported right in your inbox within 14 days!

The test is quite long and so can be challenging. Therefore, you may wish to spare at least 30 minutes in order to finish the test in one go (though it can take much longer than half an hour). You may imagine that you were taking a final examination, so that the test result truly reflects your language knowledge. It is recommended that you take the test on a laptop or desktop computer, rather than on your smartphone.

All your information will be kept strictly secret and used only for research purposes. The test will be available until June 30, 2014 or until the maximum number of test takers has been reached.

Now, let's start. First, what is your name (or nickname/pseudonym)?

Next

Figure 48 ConCloze 5 introduction page

3.4.3 Test responses

ConCloze 5 collects 285 responses altogether. This number is based on the respondents completing the first question asking their names, nicknames, or pseudonyms. The question is set mandatory for score-reporting purposes and for starting a log of test response. Illustrated in Figure 49 below, 65% of them (185 of 285) continued to at least the first item, and 34% (97 of 285) completed the entire test. The final dropout rate is thus at 66% (188 of 285).

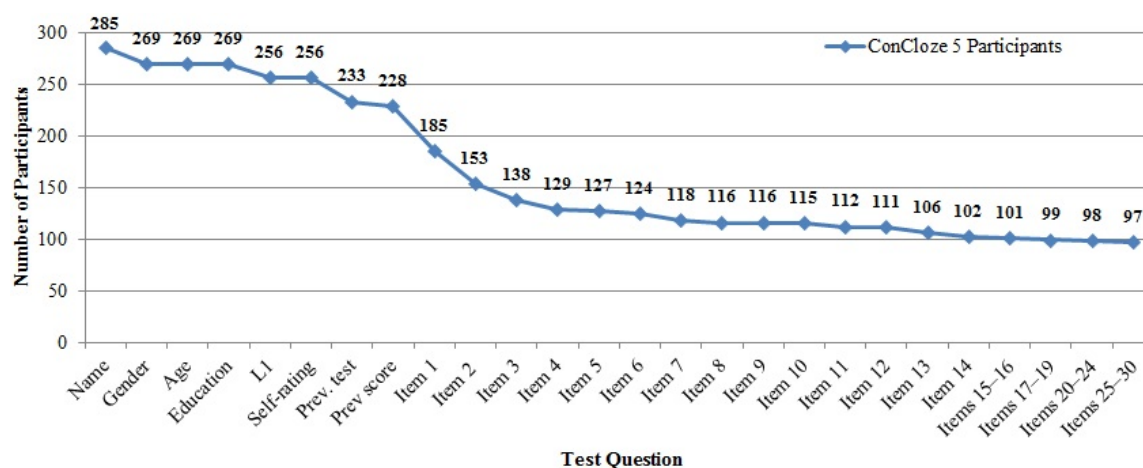


Figure 49 ConCloze 5 participant number

Superficially, the dropout rate may seem unusually high, which might suggest a high level of response invalidity in the responses. For example, ConCloze 5 might have been greatly difficult for the whole sample, so much so that a number of respondents decided in the midst of doing the test not to finish. Those who managed to finish the test might have done so in pure test inertia, a construct-irrelevant variance to score interpretation. In such cases, the scores would then not entirely reflect their true proficiency of the construct but inactiveness in task engagement.

In spite of the potential response invalidity, the following is some evidence that, when considered collectively, may dispute the likelihood of such invalidity. First, Figure 49 above depicts a gradual decline in respondents as the test progresses. However, the decline seems sharp when the number of respondents falls from 228 (previous standardized test score) to 185 (sample item—Item 1), a fall of nearly 20%. In other words, this decline occurs when (a) the sample item—not requiring responses and hence no

records shown thereof in Figure 49—and (b) Item 1 are first presented to the examinees. Since ConCloze is advertised to prospective respondents as a new test format (cf. Figure 48, page 156), this sharp decline may indicate a drop of interest. Namely, once some of the respondents had learned of the format, they might no longer feel inclined to continue and so dropped out of the test. In other words, within the first few questions, their attention to innovativeness of the test waned; thus, the participation rate becomes halved in size of the initial respondents. This would indicate that test difficulty might not be a prime factor for their decision to leave.

Apart from the psychological explanation, the high dropout in Figure 49 above may also be attributed to the stakes in testing. First of all, the rate is not unique; it is similar to the counterpart in ConCloze 1 (page 86). ConCloze has low stakes to the respondents, and so some of them may have little attention to finishing the test. On this account, the similarity in drop-out rates might not be attributed to response invalidity, but to the limited stakes of the test, which is a test-method variance intrinsic to the very nature of this research.

In addition to a drop in interest to participate, a second piece of evidence in support of appropriate test difficulty is the average item difficulty. In ConCloze 1, an average IF of 0.46 is obtained from all the items, and the test is found to be somewhat too difficult for the sample (cf., e.g., pages 92 and 102). In the current administration (to be elaborated later on page 168), the average IF is 0.67. Considering its size, this difference from the ConCloze 1 IF may be deemed distinct, indicating that ConCloze 5 is likely to be much easier in the overall picture than ConCloze 1. Moreover, looking into how the scores are distributed also confirms this improved tendency of overall difficulty. Depicted in Figure 50 below, the peak of the distribution curve is at 20 (full score = 30). An interpretation is that there are more respondents scoring on the higher-score zone than on the lower-score zone. This implies that the majority could score higher than 15, the test mid-point. Therefore, the current level of test difficulty could be argued as likely to be more suitable for the intended population than that of ConCloze 1; response invalidity is unlikely to be a serious construct-irrelevant threat in the current version. It is worth stating that based on demographic profiles reported by the respondents, there are four native speakers of English identified among the 97 test completers. As they are not part of the target population, for rigorous results, their responses are discarded henceforth. This

results in 93 non-native respondents who completed the entire test and are already reported in Figure 50 below.

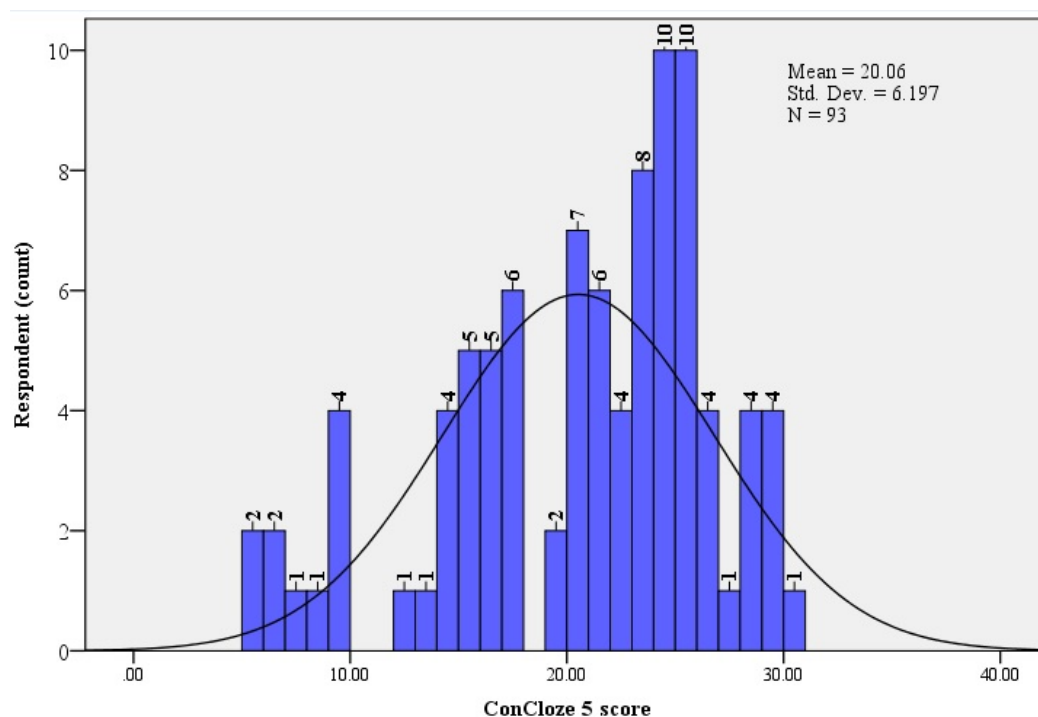


Figure 50 ConCloze 5 score distribution

The last evidence in favor of appropriateness of test difficulty is opinion of those who finished the test. This is inquired into at the end of ConCloze 5 in a feedback form similar to that of the prototyping (page 99). The form has two questions relevant to test difficulty: ‘*Question content (Comprehensible? Doable? Too easy or too difficult?)*’ and ‘*Other comments*’. The content of textual responses to these questions is analyzed, most of which can be marked with polarity in terms of test difficulty (positive–negative). The feedback is summarized in Figure 51.

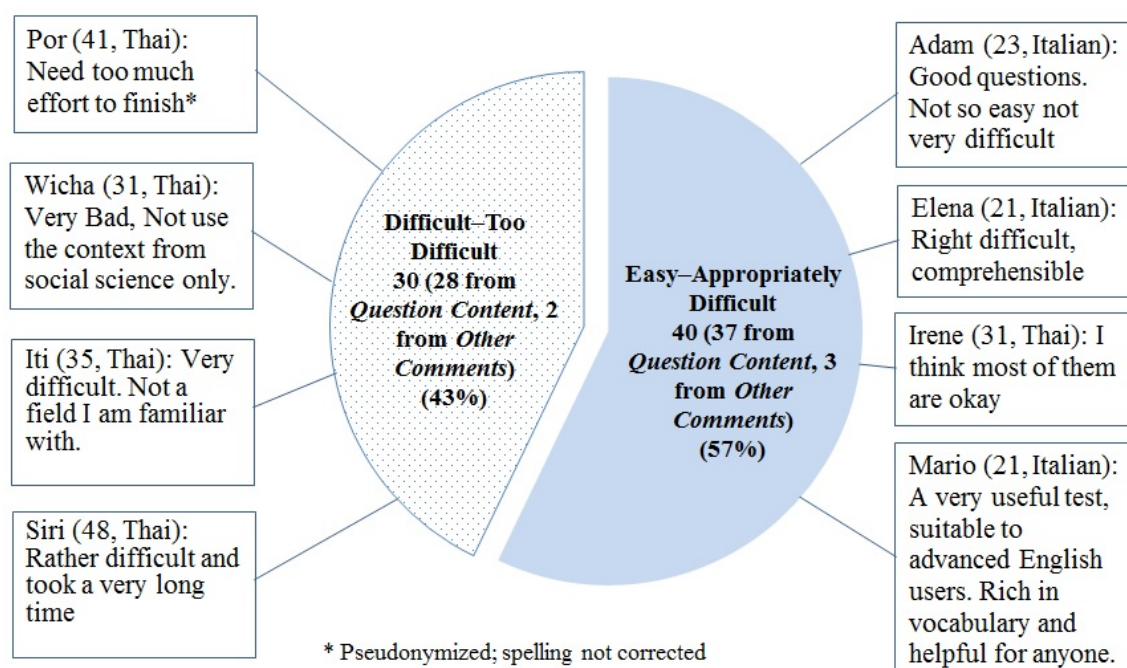


Figure 51 Respondents' feedback on suitability of ConCloze 5 difficulty

In ConCloze 1, the majority of the respondents expressed that the test was quite difficult for them (page 102). In Figure 51 above, the majority (57%) of the respondents who chose to express their opinion regarding test difficulty thought positively of the current level. This suggests that the proportion of those who found the test inappropriately difficult decreases. An inference is that the level of test difficulty, again, is generally improved from that of ConCloze 1 and could better match the average level of proficiency of the intended population.

The evidence presented thus far consists of (a) a similarity in the respondents' dropout, (b) an improved average IF, (c) a normal score distribution with a high mean, and (d) the respondents' positive feedback. These pieces of evidence allow an interpretation that the item responses from the test completers are unlikely to be much tainted by response invalidity, if any. On this account, it may be argued that the scores obtained can appropriately represent the ConCloze construct proficiency.

Given that the issue of response invalidity has been dealt with, the investigation will proceed into scale-level and item-level analyses. Concerning scale-level analyses, as introduced in the rationale (page 148), the inquiry focuses on the relations between ConCloze 5 scores and demographic variables. To begin with, correlations between the

scores and numerical variables are first explored, which are both reported in Table 27 below. It is worth pointing out that in order to represent the data with a great effect size, partial-test responses are incorporated where possible. Also, for convenience, ordinal variables (education levels and self-ratings) are treated as numerical ones.

Table 27 Pairwise correlations between ConCloze and numerical demographic variables

	ConCloze 5 Score	Age	Edu*	Self-rated Speaking†	Self-rated Listening†	Self- rated Reading†	Self-rated Writing†
Mean	20.06	29.88	5.35	2.55	2.74	3.05	2.62
SD	6.20	6.49	1.87	0.81	0.77	0.74	0.77
N	93	177	177	178	178	177	178
ConCloze 5 Score	Pearson Correlation	1.00	−0.13	−0.03	0.24‡	0.29§	0.34§
	Sig.	.	.224	.769	.019	.005	.001

* Category 1 'Now studying in a preessional course to an undergraduate level', 2 'Now studying in an undergraduate program', 3 'Already hold a bachelor's degree', 4 'Now in a preessional course to a postgraduate program', 5 'Now studying in a taught postgraduate program', 6 'Already hold a master's degree', 7 'Now studying in a doctoral program', and 8 'Already hold a Ph.D. degree'

† Category 1 'Poor', 2 'Fair', 3 'Good', and 4 'Very good'

‡ Significant at 0.05 level

§ Significant at 0.01 level

In Table 27, ConCloze 5 scores are found to have no significant association with the age of respondents ($r = -0.13$, $n = 93$, $p = 0.224$). An inference is that as the respondents grow older, it is not necessarily that they will perform proportionately better in ConCloze tasks. Lack of significant association is also true with the variable of education levels ($r = -0.03$, $n = 93$, $p = 0.769$): as the respondents are in a higher level of formal education, they may not necessarily score higher in ConCloze.

Notwithstanding the non-significant associations with age and education levels, positive correlations are discovered between the scores and self-ratings of all discrete skills in English (e.g., self-rated reading $r = 0.32$, $n = 93$, $p = 0.002$). Generally, this finding would indicate that the greater a respondent perceives themselves to be in terms of English skills, the more likely they will also perform well in the test tasks. However, caution needs to be exercised in interpreting the statistics because their effect sizes may be considered limited. For example, the coefficient between the scores and the self-rated reading—the highest among the four skills—is 0.34. Their shared variance is thus merely 0.12 ($r^2 = 0.34^2$). On the one hand, an implication is that performing well in ConCloze is most deeply related to self-perception in English reading, an interpretation which could

be meaningful when compared with the lower coefficients related to the other three skills. Yet, only 12% of the score variance can be explained by the variability in this self-rating of the reading skill, suggesting it is a mediocre score predictor. Accordingly, an argument could be that the predictive power of all the demographic variables in Table 27 is of negligible size, and a greater portion of score variance still remains unaccounted for. To verify this proposition, residuals from a stepwise regression model with these variables are illustrated in Figure 52.

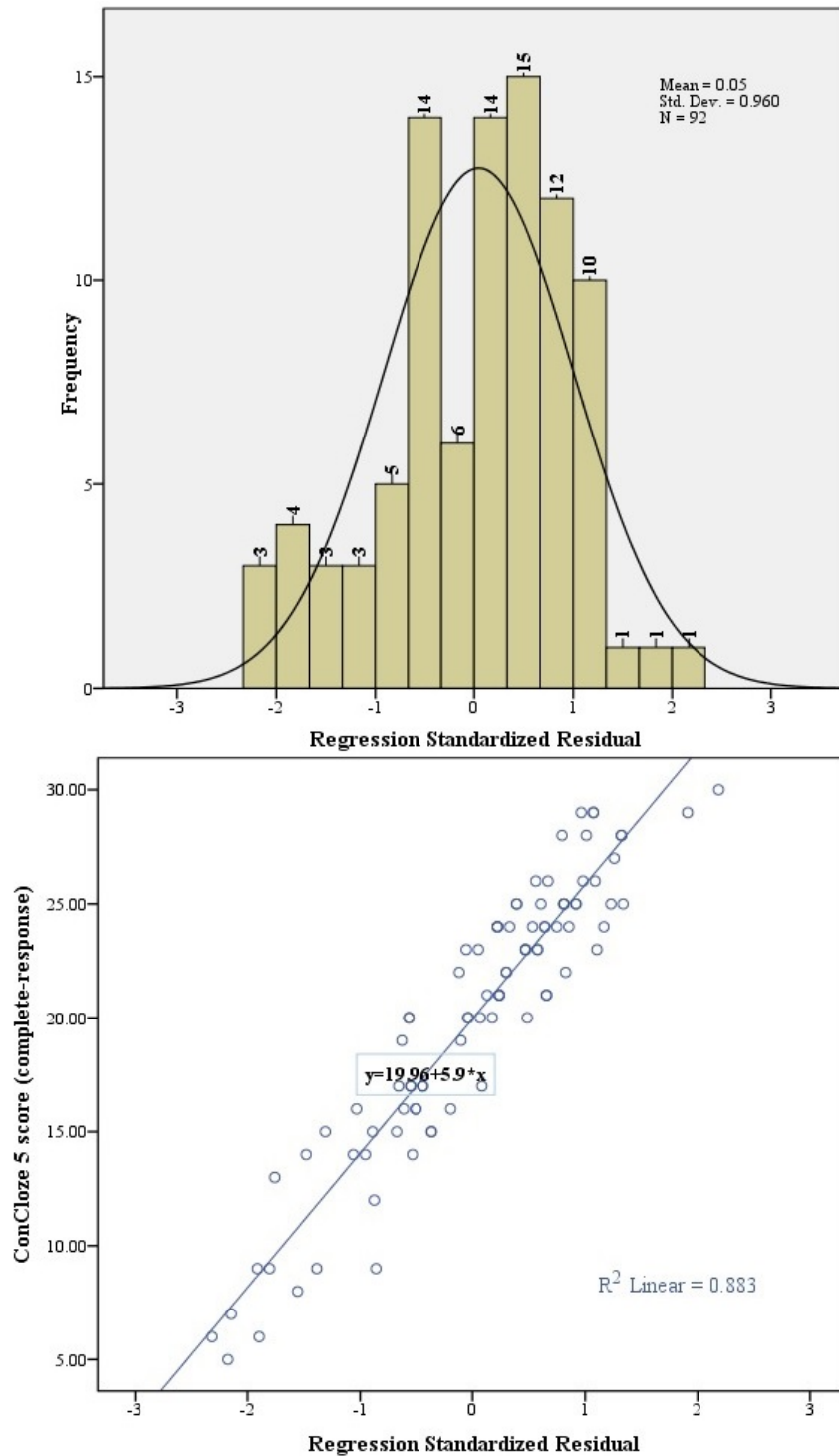


Figure 52 ConCloze 5 and regression standardized residuals

A usual assumption for a regression analysis is that non-predictor variances are unsystematic and random. This is in the sense that when most of the variances are identified and explained by a certain set of predictor variables, the remaining variances belong to no major variable. In such cases, the remaining variances could be left virtually

randomly scattered and insignificant, a situation which may be demonstrated by the equation, *Observed variance – Predictor variance = Error variance*. However, in Figure 52 above, residuals from the regression modeling appear systematic. Namely, in the histogram, approximately two-thirds of these standardized residuals vary within a range of one standard deviation from their average. An interpretation for this is that the residuals are distributed systematically in a normal curve. Because being normally distributed is a characteristic of observations attributed to independent variables, this histogram may offer evidence in support of an existing independent variable underlying the regression residuals.

In addition to their systematic distribution, the residuals can also be explained by a linear trendline of the scatterplot in Figure 52 above. This finding suggests that there is an implicit predictor variable outside of those demographic variables from Table 27 (page 161). In fact, assuming there being only one predictor variable that underlies these residuals implicitly, the predictor variable may explain up to 88% of the score variability ($r^2 = 0.88$). Therefore, Figure 52 seems to support a case of a prime predictor in the test responses that is extraneous to the numerical demographic variables and can account for most of their variability. The likelihood is that this predictor is actually the construct domain governing the responses in the ConCloze format.

Towards the validity argument, the evidence outlined above is important for three reasons. First, ConCloze 1 provides descriptive evidence of internal consistency, and ConCloze 2–4 substantive evidence of consistency in test-task content and verbalized processes. Hence, Table 27 (page 161) and Figure 52 offer the first inferential evidence that indicates the predominance of a construct proficiency in responses to ConCloze. As argued in the rationale (page 147), it adds a new facet of construct-validity evidence to this research, which, as a second reason, serves as a confirmation to those findings in the prototyping phase. Namely, all the evidence from ConCloze 1 to ConCloze 5 points to a core domain accounting for the variability of ConCloze scores. As such, the consistency across several phases of this validity study may give more confidence in drawing inferences about the construct domain, which is especially meaningful considering a larger sample size in this version than those of all the previous ConCloze versions combined. Lastly, as also hypothesized in the rationale (page 148), this evidence is the first in this study to rebut uncertainty as to relationship between the scores and demographic variables. In validation terms, this means that a source of construct-

irrelevant threat is being dealt with empirically; the validity argument is strengthened as a result.

In addition to indicating the predominance of a core construct, the correlation analyses in Table 27 (page 161) may also help with construct definition. Already reported, all the self-ratings correlate significantly positively, albeit weakly, with ConCloze 5 scores. A theoretical postulate for this pattern is that the self-ratings and scores co-vary positively not because the self-ratings cause score variance directly. Rather, the likelihood could be that they are likely to share the same source of variance and consequently correlate positively (cf. Kline 1991: 5 for ways of interpreting correlations).

For clarification, it is worth restating that prior to this part (3.4), the ConCloze domain defined is lexical-semantic knowledge mobilized primarily through the process *Testing compatibility of a given word in context*. In the broadest sense, the self-ratings and the ConCloze construct are both related to the English language. That is, the self-ratings represent self-perception of English proficiency level in each corresponding skill. In comparison, engaging in ConCloze is found to involve knowing the meaning of English words, a known essential part of general English proficiency (cf. Lewis 2000 for a central role of vocabulary in English use). In this way, for example, a strong respondent would rate themselves to be very good in all areas of English skills and, in all likelihood, also score high in ConCloze. Given the reasons outlined thus far, the self-ratings could be deemed confounding factors to the response modeling: they could be empirical surrogates of English proficiency, upon which task performance under the ConCloze construct also depends. For validity investigation, a description can be as follows. In ConCloze, the examinees are tested on knowing word meaning. This knowledge contributes positively to self-evaluation of all the four discrete skills in English. The knowledge is activated when the examinees need to figure out if one option goes well with the concordance.

In addition to associating with numerical variables, correlation of the scores with categorical demographic variables will also be investigated. Effects of two variables, gender and L1, are modeled in Table 28 below. For robust results, responses in categories with $n \leq 3$ —e.g., Category 3 of the gender variable ‘Prefer not to answer’—are excluded listwise.

Table 28 Linear modeling between scores and categorical demographic variables

Source*	Type III Sum of Squares	df	Mean Square	F	Sig.	η_p^2
Corrected Model	367.97†	5	73.59	2.09	.075	0.11
Intercept	6104.31	1	6104.31	173.05	.000	0.68
L1‡	32.02	2	16.01	0.45	.637	0.01
Gender§	0.43	1	0.43	0.01	.912	0.00
L1 * Gender	80.42	2	40.21	1.14	.325	0.03
Error	2857.3	81	35.28			
Total	37706	87				
Corrected Total	3225.26	86				

* Dependent variable: ConCloze 5 complete-response scores (87 in number)

† R squared = 0.11 (adjusted r squared = 0.06)

‡ Category 1 'Italian', 2 'Malay', and 3 'Thai'

§ Category 1 'male', and 2 'female'

In Table 28, changes in neither of the following significantly have a main effect over the scores: gender ($F(1, 81) = 0.01$, $p = .912$, $\eta_p^2 = 0.00$); L1 ($F(2, 81) = 0.45$, $p = .637$, $\eta_p^2 = 0.01$); or interaction between these two variables ($F(2, 81) = 1.14$, $p = .325$, $\eta_p^2 = 0.03$). An example of interpretation is that differences in the respondents' L1 do not have any significant predictive power over changes in ConCloze scores. The differences may explain only 1% of their entire variation ($\eta_p^2 = 0.01$). Therefore, the respondents' gender and L1, each and together, cannot be used for reliably predicting ConCloze scores. The entire model, which combines the effects of the two variables and of their interaction altogether, can account for 11% of the score variance (r^2 or $\eta_p^2 = 0.11$). On the one hand, this value of partial η squared may be deemed quite large in terms of effect size (general criteria: 0.01 = small, 0.06 = moderate, and 0.14 = large). Yet, its adjusted r squared is only 0.06. This means that when the model is estimated to the general population, its predictability is likely to be limited to merely 6% of the score variance. Accordingly, an inference is that these two categorical variables, even when acting together, tend to be limited in predictive power over variability in ConCloze scores, and thus negligible.

In an overall picture, the above finding about categorical demographic variables is similar to the finding about numerical variables presented earlier (pages 161ff.). Both of these two groups of demographic variables could be argued to have limited predictability over ConCloze scores. The interim hypothesis related to these correlation analyses is that demographic variables do not have any statistically significant effects over ConCloze scores (page 148). Accordingly, the hypothesis could be rejected, yet with some reserve. The reservation is that not all the demographic variables examined are

influential in the score variability. For those variables which are so, their effect sizes (r^2 or η_p^2) are still small. On this account, an argument is that examinees' demographic backgrounds tend to be only marginally construct-relevant to ConCloze performance.

With the scale-level analyses already dealt with, this section will proceed into item-level analyses. IV features and their effects on item discriminability (item discrimination, ID) and difficulty (item facility, IF) will be dealt with primarily. First of all, item statistics are presented in Table 29 below. It is worth stating that only IDs and IFs will be discussed in this section; alphas-if-item-deleted will be tackled extensively in the next section 3.4.4 because of their relevance to the argument therein.

Table 29 Facility, discrimination, and alphas-if-item-deleted of ConCloze 5 items

Item	IF	Evaluation*	ID	Evaluation†	Alpha if item deleted
1	0.86‡	Easy	0.36	Acceptable	0.88
2	0.78	Easy	0.57	Acceptable	0.87
3	0.43	Moderate	0.20	Acceptable	0.88
4	0.78	Easy	0.59	Acceptable	0.87
5	0.70	Easy	0.54	Acceptable	0.87
6	0.70	Easy	0.49	Acceptable	0.87
7	0.71	Easy	0.42	Acceptable	0.87
8	0.80	Easy	0.27	Acceptable	0.88
9	0.85	Easy	0.56	Acceptable	0.87
10	0.78	Easy	0.36	Acceptable	0.88
11	0.84	Easy	0.46	Acceptable	0.87
12	0.77	Easy	0.48	Acceptable	0.87
13	0.60	Moderate	0.19	Poor	0.88
14	0.67	Moderate	0.48	Acceptable	0.87
15	0.79	Easy	0.46	Acceptable	0.87
16	0.77	Easy	0.37	Acceptable	0.87
17	0.32	Moderate	0.24	Acceptable	0.88
18	0.83	Easy	0.41	Acceptable	0.87
19	0.53	Moderate	0.38	Acceptable	0.87
20	0.67	Moderate	0.27	Acceptable	0.88
21	0.42	Moderate	0.40	Acceptable	0.87
22	0.65	Moderate	0.53	Acceptable	0.87
23	0.29‡	Difficult	0.27	Acceptable	0.88
24	0.54	Moderate	0.40	Acceptable	0.87
25	0.64	Moderate	0.51	Acceptable	0.87
26	0.74	Easy	0.44	Acceptable	0.87
27	0.58	Moderate	0.18‡	Poor	0.88
28	0.62	Moderate	0.60‡	Acceptable	0.87
29	0.69	Moderate	0.57	Acceptable	0.87
30	0.76	Easy	0.48	Acceptable	0.87
Average	0.67		0.42		0.87

* Criteria: 0–0.29 = difficult, 0.3–0.7 = moderately difficult, and 0.71–1 = easy

† Criterion for acceptable discriminability: ≥ 0.2

‡ Highest–lowest

Regarding item discriminability, 36% of all the items in ConCloze 1 (14 of 39) are classified as having poor IDs (page 96). In Table 29 above, only 7% of all the items (2 of 30) can be classified as such. This distinction may indicate that the issue of discriminability is generally much improved, allowing an inference that introducing IV features to existing items can be effective for improving test-task content and ID. Being able to improve the power of discrimination can be useful for general testing purposes because it means that the items can be reused with higher effectiveness in separating the examinees with a higher proficiency level from those with a lower one. For the present validity investigation, two key inferences arise. First, the IV features are likely to have a role to play in item functioning because when the features are introduced to ConCloze 1 items, their IDs are improved as a result. This suggests that the features may likely represent the test-task content that is processed during task engagement and accordingly give substantive-validity evidence.

A second inference from investigating item discriminability is about evaluating construct-irrelevant threats. Considering the improvement in IDs, item indiscriminability seems to have become an unlikely source of imminent construct-irrelevant threat. Given the time constraints, improving item discriminability may thus not need to be focused upon in later research stages. Also, another reason for not focusing on discriminability is that IDs can fluctuate in an inverted U-shaped manner along varied IFs (Aiken 1979). For example, an ID of 0.5 could be observed when the IF of a corresponding item is up to either 0.25 or 0.75. This means that, for example, discriminability may not be a readily comprehensible and straightforward tool for comparing IV features. Consequently, seeking discriminability for validity investigation will be dropped henceforth from the toolbox of analytical measures in this study.

Concerning item difficulty, the respondents' earlier opinions on test difficulty vary, with 57% expressing that ConCloze 5 was suitable in this respect (page 160). In Table 29 above, a similar picture seems to be portrayed: the average IF is 0.67, with a range between 0.29 and 0.86. An interpretation is that, out of all the item responses, 67% are of correct answers, and the IFs also spread well in the range of moderate difficulty. Therefore, the current finding may be deemed reflective of the aforementioned opinion on the scale level, allowing an inference that both scale-level and item-level difficulties of ConCloze 5 are likely to be appropriate.

In light of observing an appropriate level of item difficulty, its significance can be described in two following ways. First, the average IF in ConCloze 1 is 0.46 (page 94), meaning the present 0.67 for the average IF is an improvement. This is a marked distinction between the two versions, which takes place when multiple IV features are introduced to the current administration. Accordingly, these features are likely to be the causes accounting for its lowered difficulty. As planned in the spec (page 149), reducing scale difficulty by bringing in IV features seems to be an effective approach for the ConCloze item type.

The second significance of the appropriateness in item difficulty lies with the validity investigation. ConCloze 1–4 have informed that item difficulty cannot be determined by the corpus-based frequency level of the target words alone (see page 96, for example). Considering the lowered overall item difficulty, the IV features likely represent difficulty drivers. In the current rationale (page 148), it is conceptualized that there may exist patterns of difficulty effects among the IV features, and each feature may or may not affect the difficulty level unequally. Therefore, in order to offer empirical contrast among the features, the IFs from Table 29 are grouped according to their IV in Figure 53.

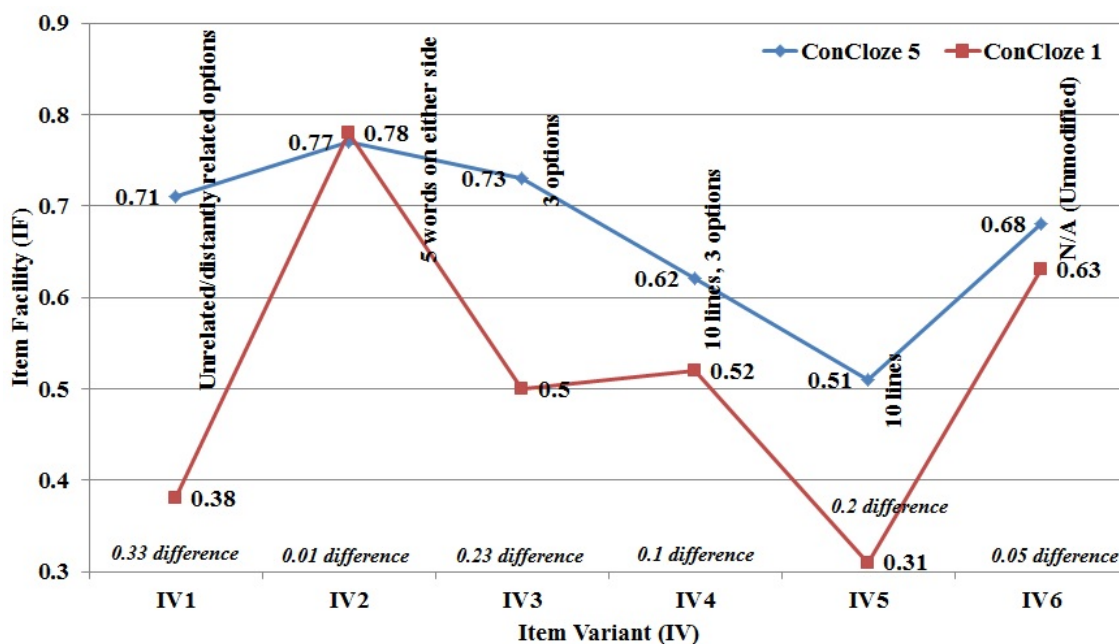


Figure 53 ConCloze 5 average IFs compared with precursor ones in ConCloze 1

From Figure 53, there could be several findings made for the validity investigation. First of all, IV6 is the control IV, representing no modification but a renumbering introduced to the precursor items from ConCloze 1 (cf. the spec, page 150). The average IF of those original items is 0.63 in ConCloze 1 whereas the counterpart is 0.68 here in IV6 of ConCloze 5. This indicates that a 0.05 difference is observed between the two administrations of these same items. Therefore, a practical inference could be that approximately ± 0.05 is a baseline variation in contrasting a given pair of item sets between the two ConCloze versions. In other words, ± 0.05 is a margin of safety in evaluating the deviation of a pair of average IFs. Also, in light of the 0.05 difference, an implication is that the two versions seem to obtain samples of the respondents with reasonably close levels of their average proficiency.

Table 26 of the spec (page 152) summarizes item statistics of the original items in ConCloze 1. It also proposes what effects to expect of the modifications to bring to those precursor items. In light of the margin of safety, significant effects of modification to ConCloze 1 items can be determined from Figure 53 above as follows. First of all, the most noticeable effect on item difficulty lies with IV1. In the spec, using unrelated or distantly related options is theoretically expected to lower item difficulty. Seemingly proved positive in IV1, using semantically unrelated or distantly related options brings about a 0.33 decrease in average difficulty. Because this is the largest difference observed among the six IVs in Figure 53, it can be inferred that using option words that are not related in terms of semantic components to one another could hugely lower item difficulty. Given this big difference, another inference is that ConCloze examinees may actively ponder upon the meanings of the option words against one another. Contrasting lexical-semantic notions can be of prime importance in ConCloze engagement.

Second to semantic relationship among the options, the next most discernible effect lies with IV5. In the spec, providing more concordance lines—from seven to ten in this case—is theoretically expected to lower item difficulty. This is also proved positive here: adding more lines is found to increase the average IF by 0.31. Therefore, it can be inferred that clues in the concordance lines could be as equally important to ConCloze examinees' processing as semantic relationship among options. Accordingly, adding more lines might mean giving the examinees more clues about the missing KWIC because this can boost their chance to score.

The next IV with IF-altering effects is IV3. In the spec, reducing the number of options is theoretically expected to lower item difficulty. This, again, is proved positive here: discarding one distractor brings about a 0.23 decrease in average difficulty in Figure 53 above. Accordingly, it can be inferred that ConCloze examinees may actively contrast the meanings of the option words. There being one fewer option potentially means that they could do so more easily, most likely in terms of both the task-processing load and probability.

The last IV discovered with positive effects on item difficulty is IV4. In the spec, reducing the number of options and adding more concordance lines is theoretically expected to lower difficulty and heighten discriminability. This is likewise proved effective through patterning in item responses: the two modifications, together, are found to decrease the average IF by 0.1, and all the items of this IV are found to be satisfactorily discriminating (Items 16–20 in Table 29, page 167). Therefore, a similar inference to the previous inferences is that having more clues and higher probability in scoring the items may really count for completing ConCloze tasks.

However, it is worth pointing out that IV4 has both fewer options and more concordance lines. This means that it is a combination in design between IV3 and IV5. Still, it is not observed that the effects of IV3 and IV5 would be combined in IV4. Specifically, the average IF of IV4 is 0.62 whereas the counterpart from ConCloze 1 is 0.52. Accordingly, the IF difference associated with IV4 is merely 0.1, rather than being able to drastically lower the average difficulty of the precursor items from ConCloze 1. This is only half the size of the difference of IV3 reported above (i.e., 0.23), let alone contrasted with the difference of IV5 (i.e., 0.31).

In ConCloze 5, IV4 is the only IV specifically demanding one option being synonymous with another (cf. the spec, page 152). In light of its meager difference in difficulty from its precursor, synonymous options could be functioning as a moderating factor. It is worth restating that synonymous options are an element of task content engineered for ConCloze 1 items. Analyzing the distractor functioning back then reveals that synonymous distractors can sometimes be attractive (page 97). Given this, an inference is that synonymous options in IV4 may make the items retain much of their difficulty. The retention could be to the extent that attractiveness of the synonymous options is in the pathway required for three options and ten concordance lines to have full

difficulty-lowering effects. In short, because the IV4 items have synonymous options, their three options and ten concordance lines cannot lower much of the item difficulty as intended.

In addition to a potentially moderating effect of synonymous options, the very finding in relation to IV1 may also help to explain the ConCloze processing in IV4. IV1 is the only IV which explicitly refrains from using synonymous options, and has achieved the largest difficulty-lowering effect in Figure 53 (page 169). It may thus follow that when options are distant in terms of meaning relatedness, they are easy for the examinees to distinguish, resulting in an easy item—the case for IV1. By contrast, when options are close in meaning, they are difficult for the examinees to distinguish, resulting in a difficult item—the case for IV4. This contrast implies that it is the core components in meaning of given options that are actively processed during task engagement. The activity in processing could be such that when the components are partially intertwined, the item difficulty could not be lowered to a great extent.

All the item modifications discussed so far seem to have discernible effects on item difficulty. These IV features can be argued as relevant to the construct proficiency. Nonetheless, the only exception lies with IV2. In the spec, cutting off the words on the concordance lines to five words on either side has been theoretically expected to increase item difficulty (page 152). From Figure 53 (page 169), this modification appears to yield an inconclusive result. Namely, the average IF of IV2 items is 0.77 in this version whereas the counterpart from ConCloze 1 is 0.78. The difference between them is only 0.01. Taking into account the baseline variation between the two versions of approximately ± 0.05 (page 170), the actual difference attributed to IV2, therefore, seems insignificant. Accordingly, the feature of IV2 does not gain enough support for claiming any real effects. An inference could be that reducing the number of words on the concordance lines may not be an effective way of making ConCloze items more difficult.

As argued in the spec (page 149), item modifications that lead to varied performance can count as processed in test-task engagement and construct-relevant. For the validity investigation, IV2—the only IV with five words on either side—is hypothesized to be the most difficult IV in the test (page 149). On the one hand, finding no discernible effect of concordance-word reduction in IV2 has rejected that hypothesis. Yet, the same finding also implies that the far co-texts are not crucial for task completion.

In other words, the clues that are still significant for the ConCloze construct are likely to be in the vicinity, rather than in the distance, of the KWIC position.

All in all, the current investigation identifies three difficulty drivers for the ConCloze item type: semantic relationship of options (closely related–unrelated), availability of clues in the concordance prompt (scarce–abundant), and number of options (three–four). The investigation has rejected a strong effect of one potential difficulty driver: number of words on either side of the concordance lines (five/ten).

3.4.4 Decision study

The previous section 3.4.3 explores scale-level and item-level responses. This section deals with reliability, another scale-level statistic. The section is dedicated to this topic because it does not directly address a validity question in this part but will eventually contribute information to determining a minimum length of the next version, ConCloze 6. The investigation begins with the reliability index of this version, which is 0.87. In the test spec (page 152), ConCloze 5 is hypothesized to retain a reliability index of 0.7 or greater, a general criterion of sufficiently high reliability for low-stakes testing. Accordingly, the 0.87 reliability obtained may be deemed high, and the hypothesis could be accepted.

In addition to the scale reliability, the role of each item in the scale reliability is also investigated. In ConCloze 1 (page 91), alphas-if-items-deleted are investigated. The finding is that deleting any of the items does not change the reliability significantly. In this version, a similar picture is portrayed: from Table 29 (page 167), there are nine items (30%) whose deletion is found to potentially drive the scale reliability to 0.88 at most. The remaining items (21 items, 70%) are found to cause no noticeable change in reliability when deleted. Accordingly, an inference is that deleting any one item in ConCloze 5 is unlikely to deteriorate or increase the scale reliability significantly as each of the items contributes nearly equally to the measure.

It is worth stressing that ConCloze 5 may be deemed an old test with IV modifications added, which is administered to a different group of examinees. In light of this research design, the significance of the above findings about reliability is twofold. The first aspect of significance lies in the validity investigation. Observing (a) a high

reliability as the test stands and (b) a stable reliability when an item is deleted indicates high internal consistency; the items are likely to consistently tap into a similar domain of competence. Finding a high internal consistency is significant because the scale in its present form is administered for the first time. This means that amidst multiple IV features, there is still a core construct likely predominating in all the item responses and making the test consistent internally. On this account, the current consistency seems to confirm the similar findings in ConCloze 1 but gives more confidence in construct interpretation because of its larger sample size.

It is worth restating that obtaining a very low internal consistency may undermine the credibility of a validity argument. This is so because it would reflect poor quality of the test and raise doubt as to the scale efficacy in construct measurement. Accordingly, the second aspect of significance resulting from the high consistency in this section is to this validity study. The findings imply that (a) in spite of multiple IVs, a ConCloze test of 30 items seems to function well in terms of internal consistency, and (b) reducing the number of items further (< 30 items) would still be theoretically viable.

In light of the theoretical viability of a test shortening, it is decided that the current number of test items (30 items) could be lowered in the next version for practical and logistic reasons (to be discussed in Section 4.2.2 of the operational use (ConCloze 6)). In preparation for the shortening, a simulation of theoretically projected reliability is performed in Figure 54 below based on ConCloze 5 item responses. The reliability indices are computed multiple times, with responses from one fewer item included each time. ConCloze 5 reliability is found to remain over 0.7 down to when the test has only nine items left. An inference is that the test seems to be really unidimensional, so much so that approximately nine items is sufficient for retaining a high reliability index for a low-stakes testing like ConCloze. Another inference is that a minimum of approximately ten items could be an informed and relatively safe estimation for a next ConCloze version to achieve a reliability of 0.7 or higher. This estimate should likely minimize the following threats. First, given the unidimensional design, the construct domain tested might not be underrepresented by a smaller number of items. Secondly, a test shortening to a moderate length is unlikely to decrease the scale reliability greatly.

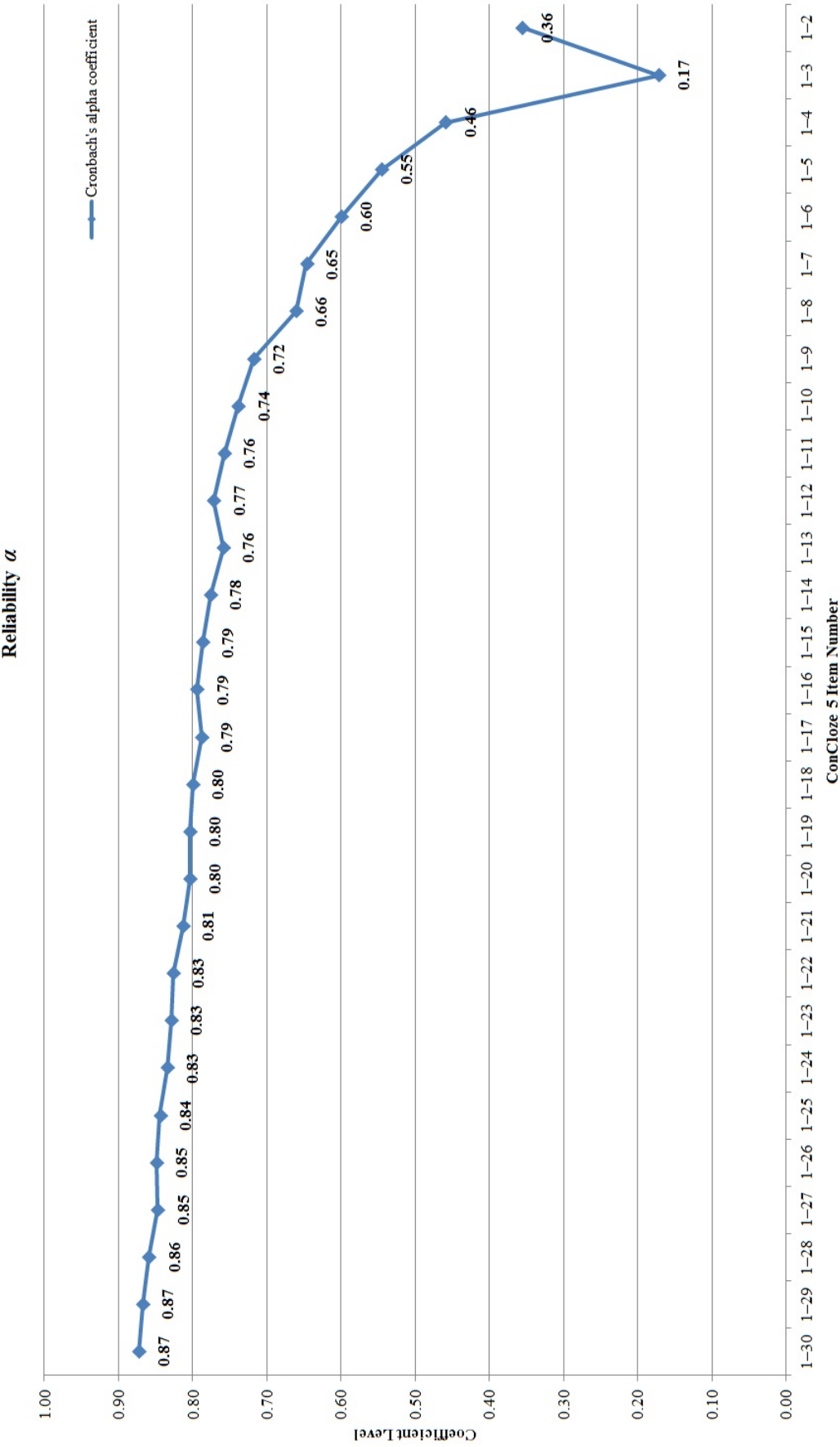


Figure 54 ConCloze 5's reliability in simulated test truncation

3.4.5 Summary

In this part, several inferences have been made based on evidence from ConCloze 5 test and item responses. They are summarized in Table 30 in no particular order.

Table 30 Major ConCloze 5 evidence and inferences

Entry	Evidence	Inference
1	1) Pattern of sharp decline in respondents across ConCloze versions 2) Much higher average IF in ConCloze 5 than in ConCloze 1 3) Respondents' positive feedback on test difficulty	Minimal response invalidity, and item responses largely reflective of construct domain
2	Lack of strong association between ConCloze scores and all demographic variables	Construct domain unlikely to be explained by variation in demographic backgrounds, but more likely explainable partly by English lexical-semantic knowledge
3	Overall improvement in item difficulty and discriminability	IV features processed in ConCloze engagement and generally construct-relevant
4	IV with semantically unrelated or distant options having the greatest effect size in lowering item difficulty	Semantic relations among the options of prime importance in the construct domain
5	IV with more concordance lines added having the second greatest effect size in lowering item difficulty	Construct-related clues embedded in the concordance lines
6	IV with one fewer option lowering item difficulty	Number of options being a difficulty driver and construct-related
7	IV with five words on either side obtaining an average IF very close to that of the precursor items	Near co-texts to the KWIC more important for task completion and more construct-related than far co-texts
8	IV with one fewer option, three more concordance lines, and two synonymous options having a mild effect on item difficulty	Synonymous options specifically invoking shared semantic components in the construct domain
9	High reliability and consistent alphas-if-items-deleted	ConCloze construct predominant in all item responses, despite all the IV features added

Considering the summary in Table 30, ConCloze 5 seems to have answered several questions surrounding the construct domain and definition. First, the potential issue of response invalidity is examined because of the high dropout rate of participants. Based on multiple sources of evidence reviewed, it is likely that response invalidity is not overwhelming but minimal. Secondly, possible relationships of ConCloze with demographic variables are also investigated, resulting in all the demographic variables

found collectively to have negligible influences over score variability. Rather, there seems to be an implicit independent variable in the analytical model that may be able to adequately account for the variability. This variable has been interpreted as the ConCloze construct domain being defined. Moreover, a pattern of correlations between the scores and self-ratings also emerge, which could be explained by their sharing a similar cause of variance, English vocabulary knowledge. Lastly, the topic of test reliability is also explored. It is found that the ConCloze 5 scale possesses high and stable reliability, which indicates that a primary construct exists that governs the measure. A theoretical projection of reliability is also simulated in case of test shortening. For an optimal reliability, at least ten items is recommended for a unidimensional ConCloze.

ConCloze 5 is engineered on the concept of item modifications added to a majority of ConCloze 1 items. As important as the above investigations for construct definition is a probe into effects of IV features on item discriminability and difficulty, which is a study into substantive aspects of the construct through the lens of quantitative item responses. It is found that semantic relationship among options is just as significant for task completion as potential availability of the clues in the concordance prompt. The number of options is then factored in as giving more or less information in arriving at a correct answer. The number is hence increasing a higher or lower probability for scoring in the item type. As to the construct evaluation in light of the test purpose, it will be discussed in the following part (3.5) because it accords with the overall evaluation of the test-development process.

3.5 Concluding Inferences and Decision

In this chapter, test and item responses to a ConCloze format have been analyzed using an exploratory approach. The following is a summary of the findings, and will be followed by an appraisal of the hypothetical construct in light of the test purpose set earlier in Chapter 1 (cf. pages 22–23 for the line of inquiry). In ConCloze 1, univariant items are generated. The responses to these items achieve a high internal consistency, suggesting a possibility that the items measure the same domain of competence. Further, their IFs spread widely but do not vary systematically according to the corpus-based frequency level of the target words. This finding has been attributed to either the interaction between the construct proficiency and varied elements of test-task content or

the narrow frequency bandwidth of the target words. In sum, the findings in ConCloze 1 constitute initial evidence for the possibility that the item type has the potential to test a distinct area of language competence.

Then in ConCloze 2–4, a test of multiple IVs undergoes usability testing. Several construct-irrelevant issues are identified and mitigated through rapid iterations. In addition, both content-validity and substantive-validity evidence is also collected through a focused set of verbal reports. It is found that the respondents were likely to process the following test-task content: concordance prompt, options, and meanings of option words. A key process constructed is *Testing compatibility of a given word in context*, which can account for the majority of the verbalization segments identified.

In ConCloze 5, a field-test version frees the construct definition from multiple demographic variables such as education level and age. The test has six IVs in total but still achieves a high internal consistency. This seems to reflect one of the findings that ConCloze 1–4 have imparted earlier: albeit never developed and systematically validated, the item type is likely to test a distinguishable competence. On this account, the findings from ConCloze 1 up to ConCloze 5 may be deemed a confirmation of the existence of the construct competence through a multivariant, mixed-methods approach to validation (univariant quantitative ConCloze 1, multivariant qualitative ConCloze 2–4, and multivariant quantitative ConCloze 5). Up to this stage of research, the investigation has gained evidence in support of the domain of lexical-semantic knowledge, which appears recurrent throughout the test-development process and tends to be invoked by the clues in the concordance prompt.

In addition to exploring the legitimacy of a distinct construct as belonging to a new item type, investigating changes in IFs also reveals construct-related mechanisms. This could be particularly evident when IV features are introduced in ConCloze 5 and a hierarchy of difficulty effects on test tasks is observed. Figure 55 below captures this finding, in which a prime effect is assigned to semantic relations among the options. Deciding from the meaning of one option weighed against those of the others is likely to be a primary process because the item responses to an IV with semantically distant options (IV1) reflect a drop in difficulty. The drop is the most considerable when contrasted with those to the rest of the IVs, which all have semantically related options. It is likely that the examinees can differentiate the option meanings most easily when their

meanings are distantly related or unrelated. Accordingly, such ability is also reflected back in ConCloze 2–4, where a core process in task engagement is identified as testing compatibility of a given word in a concordance-based context.

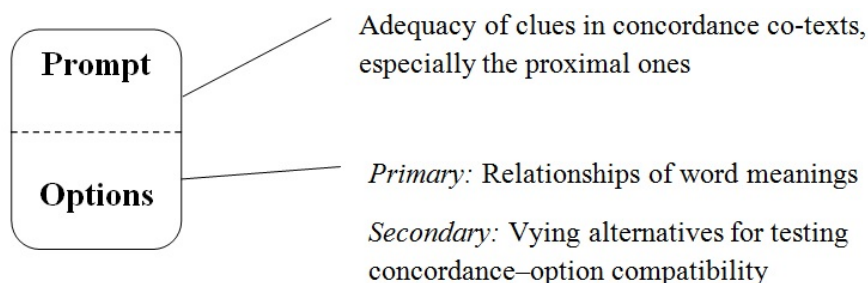


Figure 55 Difficulty effects on test-task content of ConCloze

In contrast to distantly related options, options with close meanings have been found to likely serve as a moderating factor in item-difficulty manipulation. When some of the options share core elements in meaning, item difficulty may not drop considerably, thereby lessening the usual effects of adding more clues in the prompt and of reducing the number of options. Most importantly, individual differences in item difficulty between ConCloze versions (e.g., IV1 items in ConCloze 5 vs. their precursor items from ConCloze 1) seem to likewise offer substantive insights into construct-related mechanisms. Collectively, the marked differences suggest that these IF changes are a result of IV-based modifications. Therefore, Research Question 5 in Table 1 (page 16) could be answered positively: item difficulty is likely to be affected by variation in test-task content. The main research questions that are answered in this chapter are summarized in Table 31, which also shows that Questions 6 and 7 need to be answered in subsequent research stages.

Table 31 Evaluating the operational research questions

No.	Question	Aspect	Answer
1	Are item responses internally consistent?	Structural	Yes
2	What domain(s) do the task engagements involve?	Content	At least lexical-semantic knowledge and knowledge of word association
3	What item component(s) do the examinees use in task engagement?	Substantive	Primarily the prompt and options
4	What process(es) do the task engagements involve?	Substantive	Primarily <i>Testing compatibility of a given word in context</i>
5	Is item difficulty affected by variation in task content?	Substantive	Yes
6	Is there consistency in item responses and processes across occasions?	Generalizability	Potentially. Thus far, identical item components are observed as used across items by all the sampled respondents.
7	Are ConCloze scores significantly associated with Read's (1998) Word Associates Format (WAF) scores?	External	Unanswered

In Chapter 1, the purpose of ConCloze is hypothetically set to be a proficiency test on professional and academic English grammatical and vocabulary use (page 23). In light of the construct-related findings summarized in the previous paragraphs, the test purpose can be appraised as follows. The proficiency in vocabulary use could likely incorporate lexical-semantic knowledge. As options have to be contrasted with one another using the information provided in the concordance prompt, it may be inferred that a proficient examinee could be theoretically expected to recognize the associations of a target word embedded in the co-texts as well. In sum, not only core elements in meaning of the option words, but also their associations that usually co-occur with them are tested as part of the proficiency in vocabulary use. Notwithstanding these dimensions found likely to be tested, no distinguishable item responses can yet be observed as to the dimension of grammatical use from the hypothetical test purpose.

In Chapter 1 (page 2), two elements are also referred to for evaluating validity: adequacy and appropriateness of the evidence. With regard to adequacy, this study views it as the extent to which an inference can be upheld by a particular type of evidence. Evaluating adequacy of the evidence can be important because it determines the success likelihood of a testing project amidst limited resources (cf. Fulcher & Davidson 2007; Nissan & Schedl 2012; Kenyon & MacGregor 2012). Multiple types of evidence are accrued in this chapter. In order to examine their adequacy, the evidence is categorized

according to types of investigation in Table 32 below in no particular order. The present criterion used for determining adequacy is driven by probability, looking for congruence in a facet between test administrations or congruence between facets by different test methods. The idea is that when a similar finding is observed on a second occasion, the chance is that the finding is more likely to be true.

Table 32 Major evidence in pursuit of adequacy

Investigation	ConCloze 1	ConCloze 2–4	ConCloze 5	Evaluation
Task Administration Conditions	✓	✓		Adequate
Score Distribution			✓	Inadequate
Substantive Processing		✓		Inadequate
Internal Consistency	✓		✓	Adequate
Item-component Processing		✓	✓	Adequate
Demographic Correlation			✓	Inadequate
Item Statistics	✓		✓	Adequate
Test-task Content	✓	✓	✓	Adequate

An example from Table 32 is about processing item components. In ConCloze 2–4, verbalizations are investigated, finding that no respondent can finish the test task without using the prompt and the options (page 128). The effects of changes in item components on item difficulty are investigated in ConCloze 5, finding that modifications of the prompt and/or options do have effects on IFs. Hence, this latter investigation is also on processing the item components. In light of the congruence in this respect between the qualitative prototyping ConCloze 2–4 and the field-test ConCloze 5, the evidence is deemed adequate.

Nonetheless, Table 32 above also identifies types of investigation whose evidence cannot yet be considered adequate. For example, lacks of strong association between ConCloze scores and demographic variables are unveiled in ConCloze 5. The evidence is as yet a first of its kind in this study. For this reason, evidence supporting the inference of lacking strong influences from the examinees' demographic backgrounds over ConCloze scores is regarded as inadequate; more evidence would be needed in the next chapter for a decision of adequacy in this respect, for instance.

In this chapter, internal consistency or, in a more general term, reliability is also investigated. The measure is explored in ConCloze 1 and 5 as evidence of a competence underlying the item responses. Considering the pioneering nature of this testing project, the two versions obtain relatively high reliability indices (0.84 and 0.87 respectively). On the one hand, this measure could potentially be heightened further—a very high reliability, for example, can be useful for high-stakes testing programs such as TOEFL. Yet, it usually depends on the objective of an assessment program to underscore or not to underscore reliability of the scale (Moss 1994). This ConCloze research is low-stakes for the respondents (page 92), so it seems unnecessary to aim for achieving a very high reliability index. Table 32 above also finds this type of investigation has garnered adequate evidence. On this account, it can be maintained that the validity inquiry will not shift towards an attempt to attain a very high reliability index as high-stakes testing programs in general would. A passing note would be that ConCloze can potentially be fine-tuned further in order to be a scale of very high reliability.

In addition to the adequacy of evidence, the other element of validity needing appraisal is appropriateness of evidence. In this study, appropriateness contributes to the plausibility with which an inference can be upheld by the evidence. Multiple cases of plausibility are made in the current test development, which are appraised in Table 33 below in no particular order. The criterion used for determining the appropriateness is plausibility of an item design or reasonableness of an action taken in the course of testing. This is driven by justification—when a decision in the testing research is informed by reason(s).

Table 33 Major cases in pursuit of appropriateness

Area of Appropriateness	Action–Evidence	Inference
Construct Definition	Behaviors of item responses as theoretically expected (e.g., higher proportion of items with moderate IFs and discriminability); consistency between versions	Soundness of the construct inferences
Construct Definition	Shared language processes and strategies across respondents	Explainability of substantive processing
Test Response	Mitigating construct-irrelevant variance (e.g., fatigue from a long test)	Response validity
Test Response	Rapid iterations of testing usability issues	Usability of the delivery mechanism and of the test platform
Item Response	Improved item difficulty and respondents' feedback	Suitability of difficulty level
Item Response	Justifying the size of the concordance prompt and line	Likelihood of item success
Generalization	Construct-oriented test iterations	Generalizability to the universe of admissible observations
Generalization	Turning items with moderate IFs and discriminability into IVs for substantive investigation and field-testing	Representativeness of performance in verbalizations
Extrapolation	Using middle-frequency AVL	Utilizability of the target words
Extrapolation	Focused bandwidth of corpus-based frequency for the target words	Content fidelity of middle-frequency vocabulary use

An example from Table 33 above is of language processes and strategies in the second row. In ConCloze 2–4, the respondents are sampled from varied backgrounds in the intended population and invited individually. They tackled unseen test tasks and are monitored alongside throughout the verbalization sessions. They all used the same item components to complete the test tasks, a situation of data saturation. Given the dynamic assessment of usability issues, anecdotal evidence is that no verbatim rehearsal is detected. On top of this, the verbal reports are transcribed and analyzed solely by the researcher to maximize consistency in data analysis. With the Grounded Theory-oriented labeling, several processes and strategies are constructed from unknownness based on the verbalizations. On these accounts, there is no reason to believe that the respondents would collude in preparing for the verbalizations of their task engagement. Therefore, the very finding that their verbal reports can be explained by the same processes and strategies seems to indicate substantive validity of the task engagement, and hence appropriateness in this respect in Table 33.

In light of the adequacy and appropriateness appraisal, validity of the ConCloze item type can be evaluated. While more evidence will be needed for confirmation to a few of the investigations in Table 32 (page 181), most of them seem to attain adequate evidence. Regarding appropriateness, care is taken in constructing the validity argument. This ranges from observing behaviors with dynamic assessment to drawing construct-related inferences based on empirical evidence. An integrated evaluation is that the test interpretations as they stand are likely to be valid (cf. pages 1f. for Messick's (1989) model of construct validity). From a broad perspective of test development, the ConCloze item type seems to function as normally as other psychometric measures would. For example, the items have adjustable IFs, and the test an amenable reliability coefficient. Most importantly, the testing project should proceed to the operational stage, which will allow further testing of the hypotheses formulated as well as further exploring the construct competence.

Chapter 4 Using and Applying ConCloze

4.1 Introduction

In Chapter 3, the process of developing ConCloze has been discussed. It involves the first five versions, ConCloze 1–5, which reflect different stages of the validity inquiry. In this chapter, investigating the validity of score interpretations continues with efforts to address new aspects of ConCloze validity and to confirm some of the findings in Chapter 3. As part of the validity argument, this chapter is outlined in Table 34 below, where ConCloze 1–5 become empirical backing for it. The prior versions can be deemed backing for the validity investigation because they contribute to the claim of robustness of the test format. To that extent, the scores are unlikely to be grossly plagued with construct-irrelevant variance and can accurately represent the performance of the competence tested.

Table 34 Argument for ConCloze validity

Chapter and Thesis Stage	Component in Validity Argument	
1. Introduction	Research grounds	
2. Literature Review	Theoretical backing	
3. Test Development	Empirical grounds and backing	
	Evidence and inferences	
4. Test Use and Applications	Test Version	Test Length
	ConCloze 6 Operational Use	24 items long
	ConCloze 7 Substantive Fine-tuning	5, 1, and 2 items long (in presentation order)

The organization of this chapter is as follows. ConCloze 1–5 have involved a selected-response format of the item type. In Part 4.2, this format will be used for generating new items for ConCloze 6. A criterion test, Read's (1993; 1998) WAF, is also administered alongside for criterion-validity evidence. Then in ConCloze 7 (Part 4.3), the investigated format will be used for fine-tuning the score interpretations by applying to different facets of substantive inquiry. At the end, adequacy of the evidence will be assessed in terms of the construct aspects addressed, and a decision will be made for this investigation.

4.2 ConCloze 6: Operational Use

4.2.1 Rationale

In this part, two main inquiries are driven by findings of the earlier versions. The first is from Table 32 (page 181), in which adequacy of evidence is evaluated for different types of investigation into observable ConCloze behaviors. One of the investigations whose evidence is as yet inadequate is correlation analysis between the scores and demographic variables. Evidence in this respect will be sought from a new sample of examinees in this version because a similar finding would be a confirmation to the counterpart interpretation in the test-development process (pages 161f.). It is hypothesized that the examinees' non-linguistic backgrounds and experience lack a strong association with the scores. If the hypothesis is accepted, confidence in claiming their construct irrelevance will be lifted for the validity argument. This will mean that, as a psychometric scale, ConCloze elicits scores that accurately represent the intended domain of competence.

In ConCloze 1–5, several difficulty drivers are identified from patterns of item responses. These include the number of concordance lines, and semantic relations among the options. However, no effects of the frequency level of the target words have been found on item difficulty (cf. page 95). Albeit a central notion in corpus linguistics, the frequency level cannot yet be determined as part of the test-task content processed. On the one hand, this may be owing to a limitation inherent in the prototyping: item responses of the small number of respondents could not vary so as to reveal patterns, if any. On the other hand, the target words back then are all selected from a narrow frequency bandwidth. Intended for task-content homogeneity, this could be the very design of the test that makes intangible the effects of word frequency. For these reasons, the other main inquiry in this part is to amplify differences in frequency levels of the option words sampled. If truly significant for word accessibility and hence level of task difficulty, amplifying levels of word frequency could potentially make the underlying patterns noticeable. In doing so, it is hypothesized that options selected from different frequency bands will vary item difficulty systematically.

In Chapter 2 (page 47), Read's (1993; 1998) WAF is chosen as a criterion test. Given this, WAF is also administered in this part, where ConCloze scores are hypothesized to correlate positively with WAF scores. Collecting this evidence can be important for two reasons. First, it will be the first piece of criterion-related evidence in this study. Addressing a key research question (page 16), it offers another aspect of validity evidence for construct interpretations. The second importance is about test-score interpretations: in case of a positive correlation, at least part of the WAF-score interpretations could also be applied to ConCloze. If the relationship is found negative, then part of the interpretations can be negated out of the score meaning for ConCloze. In whichever case, this could help to fine-tune the construct definition for the ConCloze item type.

Apart from garnering the criterion-related evidence, another aspect of the construct that needs empirical evidence is the generalizability aspect. In Table 1 (page 16), the generalizability aspect refers to the extent of consistency with which item responses are applicable across occasions. In this version, items are going to be generated anew and administered to a new group of sampled respondents. Administering these items offers a new occasion to amass evidence of unidimensionality, in which these items are hypothesized to be aligned with the items from ConCloze 5. Being able to put the two sets of item responses in the same model will signify congruence in the domain they test (i.e., interscalar consistency) and give generalizability-validity evidence. This then allows a wider generalization of score interpretations to the universe of admissible observations as well as increased confidence in the interpretations.

This part is structured as follows. The next section 4.2.2 describes the test spec. Afterwards, the test and item responses are analyzed in Section 4.2.3, which will be in the order of those concerning demographic variables and then those of item statistics. Then in Section 4.2.4, relations between ConCloze and WAF scores are investigated. This part concludes in Section 4.2.5 with inferences, a construct appraisal in light of the test purpose, and a decision made based on the findings.

4.2.2 Test spec

In ConCloze 2–4, IVs with five and ten words on either side of the KWIC position are used, but no idiosyncratic issues could be observed on task engagement. This implies

that the task processing may not differ significantly on the basis of whether the task has five or ten concordance words. Given this interpretation, a new IV is launched in this part: that with seven words on either side of the KWIC position. Albeit never tested previously, this IV is presumed unlikely to cause construct-irrelevant variance that would be specific to the new number of concordance words. The spec is outlined in Entry 1 of Table 35 below, featured in all the items that are newly generated for ConCloze 6, viz. Items 1–5 and 10–24 (cf. also Entry 3 in Table 35). Adding one more IV to the validity investigation can be useful because it demonstrates variability of the task format, suggesting adjustability of the item type and eventually its potential usefulness for the testing industry. Appendix 4 (page 364) offers items of the current version.

Table 35 Guiding language for ConCloze 6

Entry	Guiding Language (Test Design)		
1	Except anchor items, the prompt of each item is made up of seven concordance lines, and seven concordance words on either side of the KWIC position.		
2	To enhance appearance of the concordance prompt, the concordance line with the paragraph marker (#) is to be avoided.		
3	Except anchor items, all the items have four options each. Their AVL-frequency level and semantic relation are as follows:		
	Item	Option frequency range	Semantics of options
	1–5	0–0.3K	Distant or unrelated
	6–9 (anchor items)	1–1.3K (target words only)	(See Table 37, page 191)
	10–14	0.7–1.1K	Distant or unrelated
	15–19	1.5–1.9K	Distant or unrelated
	20–24	2.4–2.7K	Distant or unrelated
4	The newly generated items are arranged in a descending order of frequency levels of the target words. The anchor items are arranged to their ascending item estimates (easy–difficult).		
5	To test a key research hypothesis about ConCloze’s relationship with WAF, all the target words and distractors are selected quasi-randomly from COCA’s AVL. The part of speech of all the options must be adjectives or equivalent.		
6	Thirty WAF items are placed before ConCloze 6 items in the battery.		
7	The scoring method in ConCloze remains binary. For WAF, a modified Correct–Wrong scoring method is used (cf. Schmitt <i>et al.</i> 2011: 118f. for a comparison of the original methods). In each item, (a) selecting one correct word is scored one, (b) not selecting one correct word zero, (c) selecting one wrong word minus one, and (d) not selecting one wrong word zero.		

In ConCloze 1–5, a narrow frequency bandwidth of 1–1.3K is used for the spec of the target words. An idea from the rationale (page 186) is to amplify differences in the frequency levels of the options used, which is also realized in Table 35 above. In Entry

3, four frequency bands are arbitrarily selected for the options from across AVL (AVL comprises 3,015 words altogether), and separated by at least a 0.4K range. For example, Items 1–5 feature the 0–0.3K band whereas Items 10–14 feature the 0.7–1.1K band. These two bands are 400 words apart in AVL, an interval from which no words are picked for the options. The interval is intended such that the cline of the AVL words is broken up into non-continuous frequency bands with pauses interspersed – a design which should make discernible their frequency-based effects on item difficulty, if any. The spec results in four new groups of ConCloze items representing four discrete frequency bands in AVL as listed in Table 35.

It is worth emphasizing that controlling frequency levels in this spec applies to all the options in each item. For example, in Table 36 below, the key *similar* in Item 3 ranks the 92nd in the AVL, and the distractors *individual*, *various*, and *positive* rank 112th, 120th, and 137th respectively—all from the same 0–0.3K frequency band. Such control is different from the counterpart in ConCloze 1–5, in which only the frequency levels of the target words are controlled. The aim is to address the following question adequately: If everything else is equal, does the frequency level of the options actually count for the item difficulty of the ConCloze item type? (cf. Fulcher 2003b: Chapter 4 for test specs and designing fixed task content). This design is expected to render observable the effects of word frequency, however right or wrong the responses to the test questions are.

Table 36 Target words and distractors in ConCloze 6

Item	Target Word (Key)* (nth in AVL)	Distractor 1	Distractor 2	Distractor 3
1	social (4)	important (11)	economic (29)	common (98)
2	international (53)	environmental (76)	cultural (79)	likely (87)
3	similar (92)	individual (112)	various (120)	positive (137)
4	necessary (170)	effective (158)	global (174)	present (173)
5	potential (227)	additional (241)	previous (256)	standard (326)
6	traumatic (1241)†	anxious (N/A)	conditioned (N/A)	disagreeable (N/A)
7	insufficient (1114)†	available (101)	meager (N/A)	–
8	applicable (1066)†	included (N/A)	suitable (923)	true (N/A)
9	desired (1007)†	amorous (N/A)	gained (N/A)	passionate (N/A)
10	continued (732)	everyday (736)	underlying (737)	integrated (745)
11	inherent (765)	emerging (770)	linear (777)	explicit (781)
12	rational (807)	objective (813)	representative (818)	residential (831)
13	established (927)	novel (941)	indirect (952)	viable (957)
14	competing (1049)	territorial (1060)	autonomous (1076)	analytical (1087)
15	noteworthy (1503)	sequential (1506)	attributable (1508)	disparate (1509)
16	indispensable (1520)	intended (1521)	nominal (1525)	divergent (1526)
17	fragmented (1590)	contingent (1596)	persuasive (1598)	detrimental (1601)
18	aggregate (1650)	conspicuous (1653)	observable (1663)	paramount (1675)
19	provisional (1801)	interdependent (1822)	literate (1832)	concerted (1849)
20	unsustainable (2402)	networked (2400)	directed (2407)	well-developed (2403)
21	affiliated (2446)	well-documented (2428)	centrifugal (2450)	indiscriminate (2441)
22	consequential (2541)	subsidiary (2547)	imposed (2558)	proportionate (2549)
23	piecemeal (2658)	consonant (2648)	paternalistic (2636)	mitigating (2618)
24	inadvertent (2701)	manifold (2702)	cross-national (2705)	germane (2707)

* All the keys and distractors are adjectives.

† Anchor items, taken from ConCloze 5, are arranged to their IFs (see Table 37, page 191).

‡ Gradient frequency levels of the options: 0–0.3K for Items 1–5, 0.7–1.1K for Items 10–14, 1.5–1.9K for Items 15–19, and 2.4–2.7K for Items 20–24.

Higher-frequency words are generally easier than lower-frequency words (discussed on pages 95f.). This concept is taken into account in designing Entry 4 of Table 35 (page 188), and the items are arranged accordingly—higher-frequency items to the front of the test and lower-frequency ones to the back. Reflecting an ascension of easy–difficult items, the arrangement is in line with a test-writing strategy called introducing an easy warm-up intended to “overcome psychological inertia” (Henning 1987: 50). As such, more examinees are expected to participate and proceed with ConCloze 6 than in ConCloze 1 and 5. There being as many examinees finishing the test as possible can be useful because the more examinees take the test, the more likely there is a wide spread of examinees’ communicative abilities. A broad spectrum of abilities is almost always

desired in item modeling as it would increase representativeness of the population's performance.

In addition to arranging the newly generated items by their theoretical difficulty, anchor items are also arranged on the same idea into the current test version. Anchor items refer to those functioning as links between two tests, which are here ConCloze 5 and 6. The details of the anchor items used are listed in Table 37 below; for example, Item 9 in ConCloze 5 becomes Item 6 in this version. The anchor items are arranged in the order of descending IFs, reflecting the same easy–difficult strategy as done to the other items in this version.

Table 37 ConCloze 5–6 anchor items

ConCloze 5 Item	IF	ConCloze 6 Item	Target Word (Key)	Number of Lines	Number of Words	Number of Options	Semantics of Options
9	0.85	6	<i>traumatic</i>	7	5	4	Related
18	0.83	7	<i>insufficient</i>	10	10	3	Related, one synonym
26	0.74	8	<i>applicable</i>	7	10	4	Related
6	0.70	9	<i>desired</i>	7	5	4	Related

There are two perspectives worth considering in determining the number of anchor items used. The first is to compare with the number of items deemed sufficient. According to Linacre (2012), at least three items are generally needed for linking two tests. As the current test has four anchor items, they should suffice to link ConCloze 5 and 6. A second perspective is to evaluate the number of anchor items in relation to the whole test. According to Angoff (1971, cited in Shin 2009: 2), approximately 20% of an entire set of test items are recommended as sufficient for test linking. The four anchor items account for approximately 20% of the 20 newly generated items and so should be adequate for the linking.

Synonymous options are found in ConCloze 5 to be a moderating factor in reducing item difficulty (page 171). This is in the sense that they may hinder other modifications in task content from exerting due influence on difficulty. While not all synonymous options would be so (as in the anchor items themselves), a precaution taken is to avoid them. Realized in Entry 3 of the Table 35 spec (page 188), all but anchor items

use distractors of distant or unrelated semantic components. This design is intended to single out other factors that may confound the difficulty effects by word frequency.

Concerning test length, it is worth stating that the present version has a much lower number of items than ConCloze 1 and 5. While a short test usually suffers low reliability, this test should unlikely experience so poor a reliability index as to implicate poor quality, a construct-irrelevant threat. This is so because a decision study in Section 3.4.4 (pages 175f.) has simulated that approximately only ten ConCloze items will be needed for a reliability index of at least 0.7, a level generally regarded as high. For this reason, the minimum number has already been taken into account in Entry 3 of Table 35 (page 188); the 20 items newly generated should be much larger than the simulated estimate 10.

Apart from investigating effects of word-frequency level, the relationship of ConCloze with the criterion test is also considered in the current spec. The consideration is in terms of (a) number of items, (b) word class, and (c) order of the tests. With regard to the number of test items, the WAF version used here originally has 40 items (Read 1998, cited in Cobb ca. 2011). Available at www.lex tutor.ca/tests/associates/test.html, Figure 13 (page 48) illustrates one of them. When administered alongside ConCloze, WAF could become part of a long test battery containing 64 items (WAF 40 + ConCloze 24). On the one hand, a long battery may not pose serious threats to test interpretations as long as a sufficient number of test completers are obtained for the types of data analysis required. Yet, lengthy testing may raise logistic issues such as high dropout rates and incomplete responses (cf., e.g., Figure 21 (page 86), Figure 23 (page 101)), which in turn would unnecessarily prolong the data collection. Moreover, issues of response invalidity may emerge among the responses of the test completers. For example, those who manage to finish the test may have done so with much fatigue or guessed for the most part of the test merely in order to finish it. For these reasons, WAF is arbitrarily shortened to the first 30 items.

The next consideration relating to WAF is about word class of the options. Set out in Chapter 2, each WAF item presents a target adjective, as in *favorable* of Figure 13 (page 48). Four words on the left-hand column are adjectives, and the other four on the right nouns. This means that the majority of the words, particularly the prompts, are adjectives. Considering this, only adjectives are also used for the options of all the

ConCloze items generated in this version as specified in Entry 5 of Table 35 (page 188). It is expected to increase potential comparability of the task content in the battery and legitimize association analyses.

The last consideration relating to WAF is about the order of the two tests in the battery. In Figure 13 (page 48), one WAF item is exemplified for presenting nine words to the examinees. Determined earlier, WAF has 30 items, presenting 270 words in total ($9 \text{ words} \times 30 \text{ items}$). By contrast, each ConCloze item presents approximately 102 words ($[\text{seven concordance lines} \times 14 \text{ words in each line}] + \text{four options}$), which is equivalent to a processing load of 2,040 words in total ($102 \text{ words} \times 20 \text{ items}$). Irrespective of the anchor items, ConCloze would very likely be a far heavier burden for processing than WAF. If the examinees tackled ConCloze before WAF, a consequence could be response invalidity: many examinees might feel discouraged from finishing the test or complete it without proper attentiveness. Therefore, in order to avoid test inertia and mitigate response invalidity, WAF is to be presented in the battery first, and ConCloze later. On this account, only one version of the battery, WAF–ConCloze, is produced by Entry 6 of the spec (Table 35, page 188), rather than two versions (WAF–ConCloze, and ConCloze–WAF).

In each WAF item, the examinees are required to choose four words. By contrast, they are asked to choose only one option word in each ConCloze item. Considering the test ordering designed, a light carryover effect of WAF towards ConCloze can be anticipated. Therefore, a short precaution, ‘Only one of the given choices is the correct answer,’ is added to the sample item of ConCloze. Illustrated in Figure 56 below, this is intended to emphasize the fact that only one option, rather than four, is the correct answer. Also, it is worth emphasizing that, on the one hand, it would be ideal to have all the target words in ConCloze that are identical to the words tested on in WAF. On the other hand, however, because target words in WAF are not based on COCA (cf. pages 47ff.), the target words in ConCloze remain AVL-derived (based on Davies’s (2008–) COCA) for consistency across the research project.

Testing Academic Vocabulary

You may need to maximize the web browser display to fit the question. Or alternatively, you may hold 'Ctrl' and press '+' or '-' to zoom in or zoom out, respectively, as you please.

The multiple-choice question below is a sample question. Study it carefully.

Note that Lines 1-7 below are texts taken from different sources, so they are not of the same passage. However, they need the same word to fill in their blanks. Only one of the given choices is the correct answer.

Sample question

1 potential cheaters, those who wish to _____ early and those who may need to use the
 2 A working model that is not yet _____ ed but that represents the major technical,
 3 will discuss it with them after they _____ ed. Do NOT tell them that you will be
 4 each lexical item. || Once you have _____ ed the interview, compare the results from
 6 and receive a score directly after _____ ing the test. || It is probably noticeable that
 5 A, and the curve for the non-masters _____ s at point C. There will be some non-masters
 7 deal of important information || To _____ this section, I will present an example of what
 place.

Special text, made
 of excerpts from
 different sources.
 Lines 1-7 here are
 not from the same
 place.

All the lines above miss the same word. Which of the following should be that word? ← question

A allow B finish(e) C participat(e) D study(i) ← choices

Figure 56 ConCloze 6 sample item

Apart from organizing the tests in the battery, the next topic in spec writing is the scoring methods for the two tests. Specified in Entry 7 of Table 35 (page 188), a modified Correct–Wrong method is used for scoring WAF items (cf. Schmitt *et al.* 2011: 118f. for a comparison of the original methods). In each item, (a) selecting one correct word is scored one, (b) not selecting one correct word zero, (c) selecting one wrong word minus one, and (d) not selecting one wrong word zero. This is an *ad hoc* scoring method, brought into use due to actual multiple variations found in responding to each item and the consequent complication of marking the responses. For example, a respondent may choose one to four words, in any combinations. In an extreme case, a respondent is found

to select all of the options in several items of WAF. As for ConCloze items, the scoring method remains binary. Also, it is worth stating that multiple options in ConCloze cannot be selected simultaneously because of mutual exclusiveness of the checkboxes.

4.2.3 Test responses

There are 576 respondents recorded as starting the response log at the website <https://www.surveymonkey.com/s/AcadVocabTesting>. Figure 57 below shows their number throughout the test battery. In ConCloze 1 and 5, the completion rate is at 34% (pages 86 and 157 respectively). In contrast, 247 of the current respondents completed the test battery, which is equal to a completion rate of 43%. While this completion rate is lower than half and does not account for the majority of the 576 respondents, it is higher than the counterparts in ConCloze 1 and 5. Also, it is worth considering that the current version is administered for the same duration as ConCloze 1 and 5, which is approximately one month. Despite this equal period of test administration, WAF is also administered alongside. Considering (a) the higher completion rate, (b) comparable duration of administration, and (c) WAF administration, the present completion rate could be deemed an improvement vis-à-vis those counterparts. Accordingly, it may be argued that the majority of those who completed the test battery were likely to do so actively. An implication is that response-invalidity issues such as severe fatigue and test inertia are unlikely to be pervasive in the test responses.

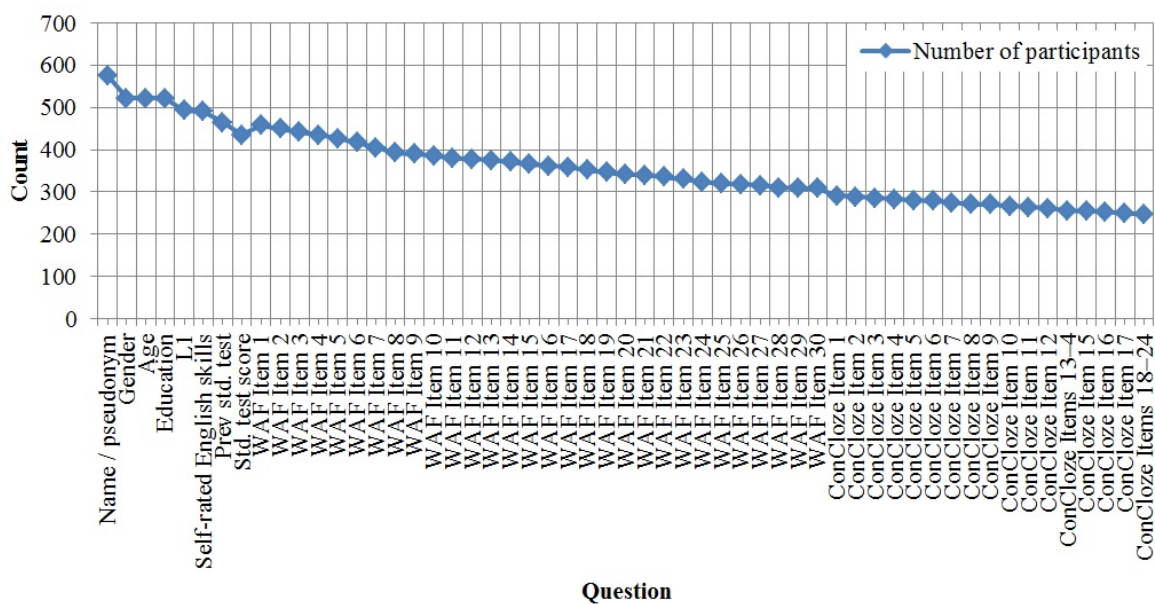
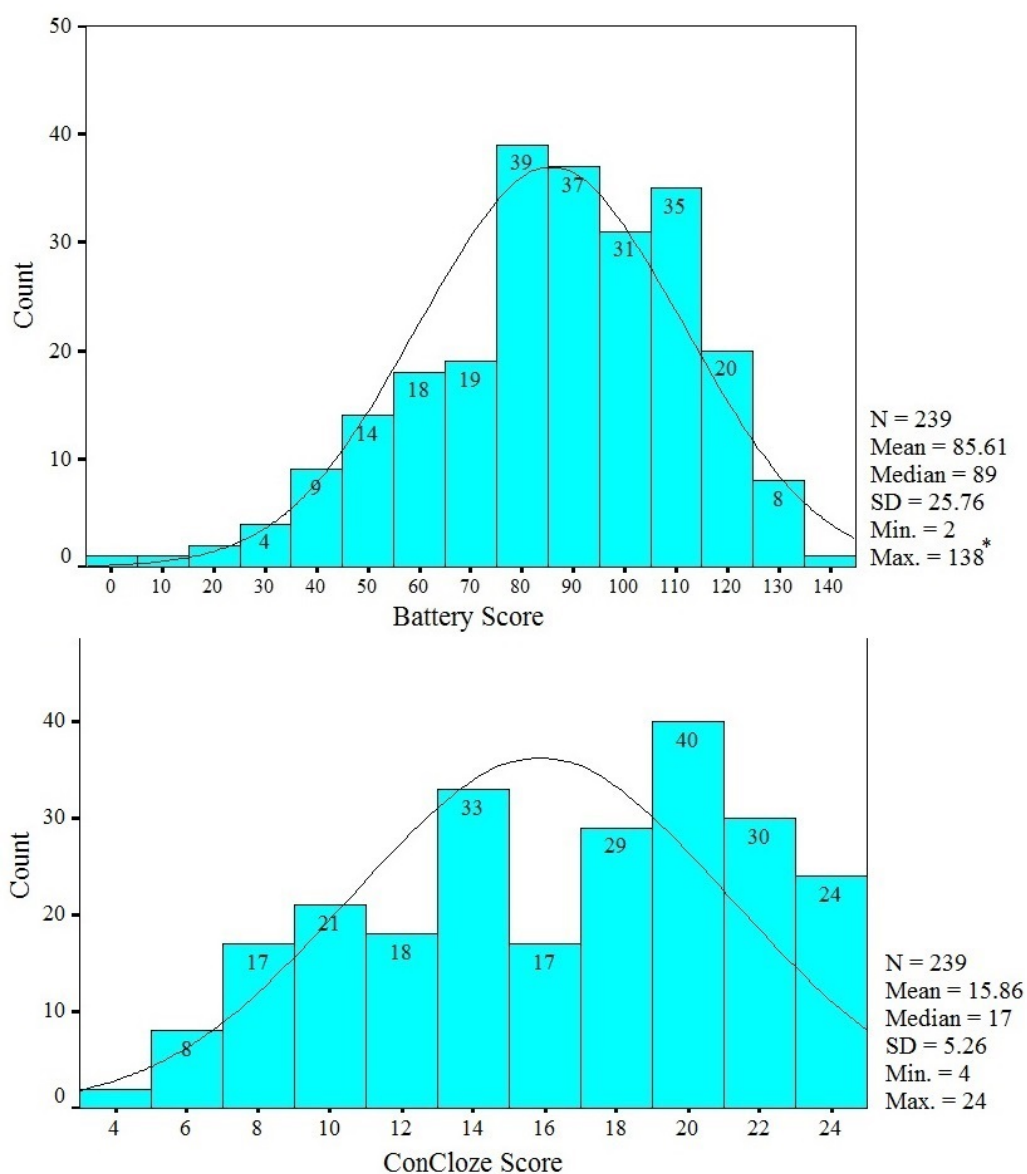


Figure 57 Number of participants across test battery

Apart from the number of respondents to ConCloze, Figure 57 above also includes the number of those to WAF. The scores in the two tests by those who finished the entire battery are added up as battery scores. The descriptive statistics and distribution of the battery scores and of ConCloze scores are displayed in Figure 58 below. The battery-score distribution is found to be slightly negative-skewed, and so is the ConCloze-score distribution. Both of their mean scores and median scores are also above the corresponding midpoints (72 for the battery, and 12 for ConCloze). An interpretation is that both the test battery and the ConCloze test seem to be moderately difficult for the sample and likely appropriate for the intended population on both the battery level and the ConCloze-scale level. This implies that, again, response-invalidity issues should unlikely be threatening to the score interpretations and, therefore, the responses to be analyzed can accurately represent the performance of the proficiency measured. It is worth pointing out that the score distribution of WAF is not displayed in Figure 58 because it is not directly related to the current construct investigation.



* Maximum possible score = 144 (120 in WAF [30 items × 4 points each] + 24 in ConCloze)

Figure 58 Score statistics and distribution of the test battery and ConCloze

In addition to the response validity, population relevance is also examined through multiple measures. First of all, the elicitation protocol is similar to that of ConCloze 1 and 5 (cf., e.g., pages 71 and 155 for details about those protocols). For example, the respondents' task completion is not time-restricted, so as to encourage their committed engagement. Another identical measure is to screen their L1 backgrounds, in which eight respondents are found to identify themselves with the English L1 and, consequently, their responses are excluded from the analyses henceforth. Considering the measures for both response validity and population relevance, it could be argued that the responses obtained

could appropriately represent the ConCloze engagement as performed by the intended population.

Given the likelihood of response validity and population relevance, the ConCloze responses will be first investigated in terms of their relationship with numerical demographic variables. In Table 38 below, their correlations are analyzed pairwise. For convenience, the ordinal variables of education levels and English self-ratings are treated as numerical ones (cf. also a similar treatment of demographic data in ConCloze 5, page 161). The scores are found to correlate significantly with all the variables, which may relate to the construct proficiency in the three following ways. First, the relationship of the scores is stronger with the education levels than with age ($r(237) = 0.24, p = .000$ vs. $r(237) = 0.13, p = .25$). An interpretation is that the respondents who are in a higher education level tend to obtain a higher score in ConCloze. By contrast, those who are old may not always score well. This pattern of correlations is identical to the counterpart in ConCloze 5 but goes unnoticed then (ibid.). This finding may suggest that the construct proficiency is developed more with education than with bodily age, and could be attributed to exposure to academic English in the education process. In other words, the ConCloze construct may take time to evolve and develop, but can grow more significantly in formal education.

Table 38 Pairwise correlations between scores and numerical demographic variables

	ConCloze 6 Score	Age	Edu*	Self-rated Speaking†	Self-rated Listening†	Self-rated Reading†	Self-rated Writing†
Mean 	15.86	28.74	4.75	2.68	2.88	3.1	2.72
SD	5.26	8.85	1.99	0.81	0.82	0.76	0.79
N	239	509	509	482	480	480	477
ConCloze 6 Score							
Pearson Correlation	1.00	0.13‡	0.24§	0.37§	0.35§	0.44§	0.42§
Sig.	.	.025	.000	.000	.000	.000	.000

* Category 1 'Now studying in a preessional course to an undergraduate level', 2 'Now studying in an undergraduate program', 3 'Already hold a bachelor's degree', 4 'Now in a preessional course to a postgraduate program', 5 'Now studying in a taught postgraduate program', 6 'Already hold a master's degree', 7 'Now studying in a doctoral program', and 8 'Already hold a Ph.D. degree'

† Category 1 'Poor', 2 'Fair', 3 'Good', and 4 'Very good'

‡ Significant at 0.05 level

§ Significant at 0.01 level

|| Both partial- and complete-test responses to the demographic variables are included in the descriptive statistics for accurate representation.

A second construct-related finding from the correlation patterns in Table 38 is about comparative effect size. The variables of education level and age are found to have smaller effect sizes than the self-ratings (*education* $r(237) = 0.24$, $p = .000$, *age* $r(237) = 0.13$, $p = .25$ vs., e.g., *self-rated speaking* $r(237) = 0.37$, $p = .000$). This pattern is also identical to the counterpart in ConCloze 5 (page 161) and, therefore, may suggest a substantive meaning for the construct. One interpretation could be that the level of the examinees' task performance is reflected more accurately by how the examinees perceive themselves regarding English skills than by their education level or age. This means that how well the respondents can perform in ConCloze tasks is systematically bound to the self-perceptions. In other words, the self-perceptions of all of the four skills could have a common factor with the ConCloze competence. Generally, self-ratings tend to load heavily on a general factor of linguistic competence (Bachman & Palmer 1989). Likewise, the ConCloze competence is also argued in ConCloze 5 for its common dependence on general English proficiency with self-ratings (page 165). Accordingly, the present positive associations between ConCloze scores and the self-ratings seem to support an argument that the construct shares the same source of variance with the self-ratings, viz. general English proficiency. Because lexical-semantic knowledge is fundamental to general proficiency in English, this evidence may also support the argument for lexical-semantic knowledge as part of the ConCloze construct. Figure 59 portrays this relationship: self-ratings serve as surrogates reflecting the general proficiency, which encompasses the ConCloze competence.

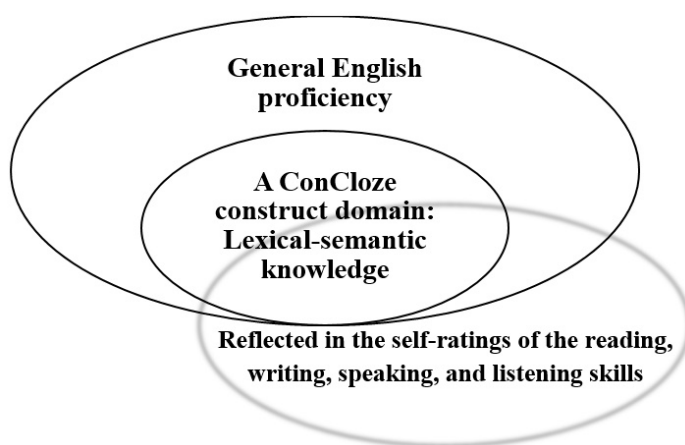


Figure 59 ConCloze construct amidst self-ratings

The last construct-related finding emerging from the correlation patterns in Table 38 is about ranking among the self-ratings. In ConCloze 5, the scores correlate the highest with the reading self-rating and the second highest with that of the writing skill. The two self-ratings are also more strongly associated with the scores than the listening and speaking self-ratings are. An interpretation is that the respondents who evaluate themselves as able in English reading tend to score high in ConCloze. This order of correlation strengths, which goes unnoticed back in ConCloze 5, is also found in this test version (*reading* $r(237) = 0.44$, $p = .000$ > *writing* $r(237) = 0.42$, $p = .000$ > *speaking* $r(237) = 0.37$, $p = .000$, *listening* $r(237) = 0.35$, $p = .000$). This congruence between the two test versions could signify its meaningfulness: the ConCloze proficiency may be most deeply related to self-perceptions in English reading and writing, respectively. An inference could be that the lexical-semantic knowledge mobilized in ConCloze engagement is more of that used in English reading and writing than in English listening and speaking. Also, it is worth pointing out that the range of vocabulary used in written language tends to be generally wider than that used in spoken language. This implies that as the ConCloze format is in written form, the range of vocabulary used could be so great as to being reflected in the association patterns between the ConCloze scores and English self-ratings.

Notwithstanding the substantive inferences derived, the correlations should be interpreted with caution. This is so because of their limited effect size; for example, the reading self-rating obtains the highest coefficient 0.44 in Table 38 (page 198), equivalent to an r squared of 0.19. It implies that, on the whole, up to 19% of the score variance could be explained by the demographic variables, and a larger portion of score variance remains unaccounted for. To verify this interpretation, the regression residuals of these variables are modeled in Figure 60.



Figure 60 ConCloze 6 and standardized regression residuals

In a regression model, residuals are expected to be unsystematic and reflect no major predictor variable (cf. also an earlier argument in ConCloze 5, pages 163f.). In Figure 60 above, the standardized residuals are ogive-shaped rather than unsystematic, a pattern which can be significant for two reasons. First, the evidence indicates a predictor variable other than the demographic variables that is central to the score variance. This pattern is identical to the counterpart in ConCloze 5 (ibid.) and seems to suggest the same substantive meaning discussed earlier regarding the self-ratings. Namely, the ConCloze construct is likely to underlie the score variance. The second reason is for generalizability evidence. ConCloze 5 and 6 are composed of different sets of items and administered separately. Finding identical patterns of systematic regression residuals indicates consistency across occasions and gives a power of generalization to the score interpretations.

In addition to the correlations with numerical variables, the scores' association with categorical variables is also investigated. Effects of two demographic variables, gender and L1, are modeled in Table 39. For robust results, responses in categories with $n \leq 3$ —e.g., the category 'Prefer not to answer' of the gender variable—are excluded listwise. It is found that changes in categories of neither of the following factors have a significant main effect on the scores: gender ($F(1, 173) = 2.2, p = .14, \eta_p^2 = 0.01$), or interaction between gender and L1 ($F(6, 173) = 1.5, p = .182, \eta_p^2 = 0.05$). It can be interpreted that, for example, differences in the respondents' gender are unlikely to have any significant predictive power over changes in ConCloze scores. The differences may

explain only 1% of their variance ($\eta_p^2 = 0.01$). Therefore, the respondents' gender and its interaction effect with their L1 cannot be used for reliably predicting ConCloze scores.

Table 39 Tests of between-subjects effects on ConCloze 6 scores

Source*	Type III Sum of Squares	df	Mean-square	F	Sig.	η_p^2
Corrected Model	1913.42†	13	147.19	7.86	.000	0.37
Intercept	17628.13	1	17628.13	941.04	.000	0.85
Gender‡	41.18	1	41.18	2.2	.14	0.01
L1§	1572.30	6	262.05	13.99	.000	0.33
Gender * L1	168.29	6	28.05	1.5	.182	0.05
Error	3240.75	173	18.73			
Total	48712	187				
Corrected Total	5154.16	186				

* Dependent variable: ConCloze 6 complete-test scores (187 in number)

† R squared = 0.37 (adjusted r squared = 0.32)

‡ Category 1 'male', and 2 'female'

§ Category 1 'Arabic', 2 'Chinese', 3 'Hindi', 4 'Italian', 5 'Telugu', 6 'Thai', and 7 'Urdu'

However, Table 39 above also shows several tests of between-subjects effects that turn out significant. Firstly, the L1 variable has a significant effect on the scores ($F(6, 173) = 13.99, p = .000, \eta_p^2 = 0.33$). The partial η squared of 0.33 could also be deemed very large in effect size (general criteria: 0.01 = small, 0.06 = moderate, and 0.14 = large). An interpretation is that differences in the respondents' L1s likely have significant predictive power over changes in ConCloze scores. Further, the entire model—combining all the factors and their interaction—has a significant main effect on the scores ($F(13, 173) = 7.86, p = .000, \eta_p^2 = 0.37$). As gender and its interaction effect with L1 do not have significant effects on the scores, it may be inferred that L1 is likely to be the primary factor driving the model to attain a high predictability over the scores. Moreover, the adjusted r squared of the model is 0.32, which means that when estimated to the population, its predictability is up to 32% of the score variance. Accordingly, the respondents' L1s can be inferred as a good predictor of the ConCloze-score variability.

In ConCloze 5 (pages 166f.), changes in L1s do not have significant effects over the scores. Given this, the current finding is in contradiction and may contain patterns related to the construct. As the variable of education level is previously found to correlate positively with the scores (pages 161 and 198), it is hypothesized that the educational variable may function as a confounding factor in the relationship between the scores and

different L1s. To investigate this, the estimated marginal mean scores and education levels of the respondents with different L1s are portrayed in Figure 61.

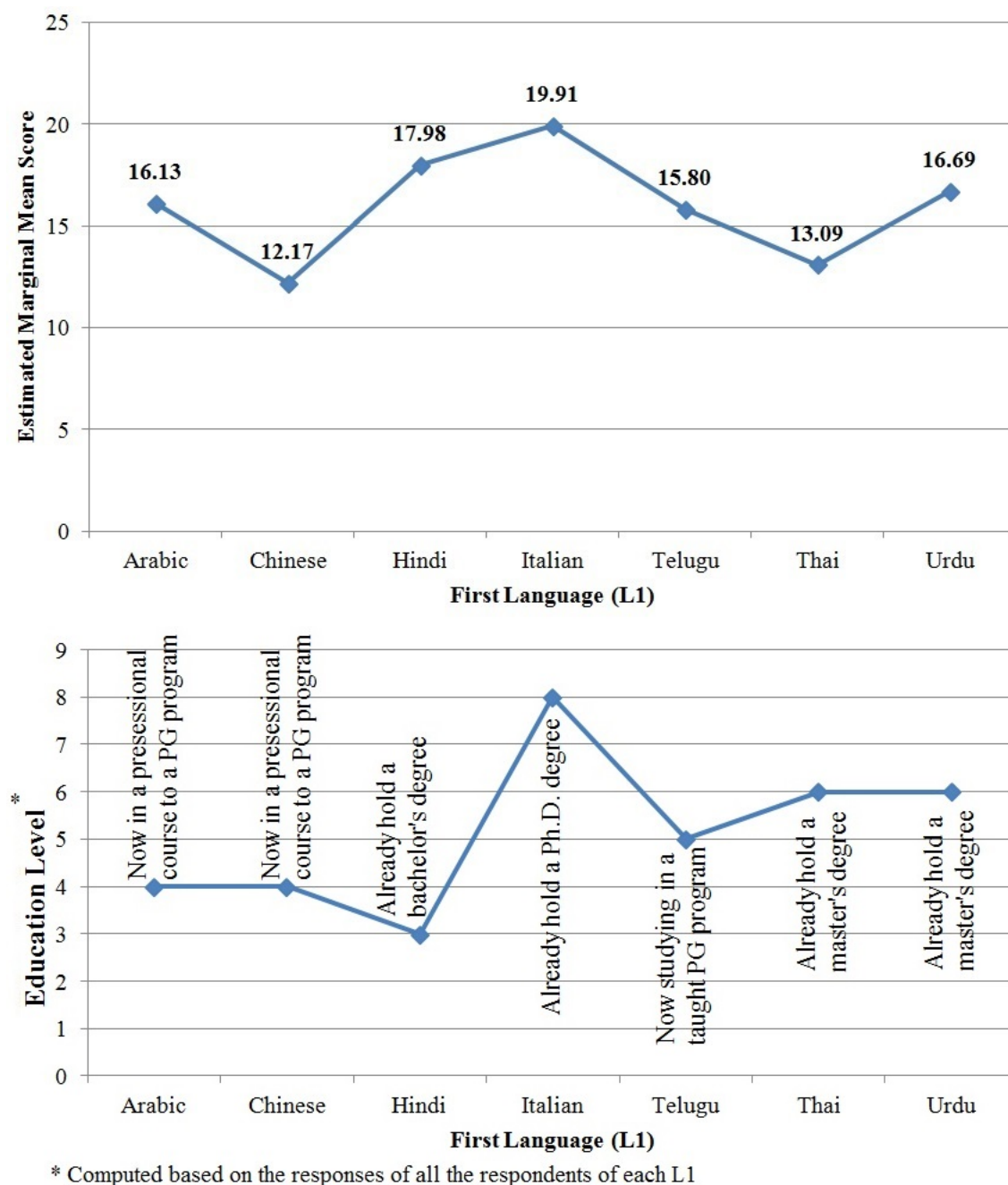


Figure 61 Estimated marginal mean score and education level in L1 groups

Language is a mirror of sociocultural contexts. In Figure 61, two contrastive pairs may demonstrate this notion and give substantive evidence: Chinese vs. Italian, and Thai vs. Hindi. In the former pair, the Chinese mode of responses in the variable of education

level is 'Now in a preessional course to a postgraduate program'. By contrast, the Italian counterpart is 'Already hold a Ph.D. degree'. An interpretation is that the largest group of Chinese examinees were still in the intermediate higher education when taking the ConCloze test. By contrast, the Italian counterparts had been through the highest stage of formal education. This difference in education has two implications. First, in terms of language learning, the Chinese may likely have had lower length and intensity of their exposure to academic English vis-à-vis the Italians. Secondly, the lower mean score in the Chinese group than the Italian counterpart seems natural as this reflects the probable difference in such exposure. On this account, L1 serves here as a surrogate of co-influentiality between the ConCloze-proficiency level and education, a social factor through which the examinees acquire the competence.

In contrast to education as a potential indirect influencer on ConCloze performance, the Thai–Hindi pair seems to portray a slightly different scenario. The Thai mode of responses in the variable of education level is 'Already hold a master's degree'. By contrast, the Hindi counterpart is 'Already hold a bachelor's degree'. An interpretation is that the largest group of Thai examinees had already been through the intermediate level of higher education when taking the ConCloze test. Their counterparts had obtained a degree of the foundational level of higher education. This contrast is the converse of the Chinese–Italian case: the Thai group has a lower mean score despite the mode of a higher level of education. However, it is worth considering that the Thai language is known to be spoken primarily in Thailand, where English is used mainly for educational and foreign affairs rather than for everyday purposes. By contrast, Hindi is typically spoken in South Asia, where English functions as a practical lingua franca and also an official language (Crystal 1997, 2003: 12, 48f.). This implies that the Thais may likely have had relatively limited length and intensity of their exposure to academic English when compared with the Hindi speakers. On this account, L1 serves as a surrogate of the extent in which the examinees had used and been exposed to academic English. It reflects co-influentiality between the ConCloze-proficiency level and sociocultural settings in which the examinees acquire the proficiency.

Considering the Chinese–Italian and Thai–Hindi findings interpreted, a collective inference is that the acquisition of the proficiency tested by ConCloze may partly depend on the intensity and extent of exposure to English. The more extensively the examinees have been in the contexts of English use, the more likely they would develop a higher

proficiency level for ConCloze performance. That said, caution needs to be exercised in assessing this construct-related inference because the current examination of L1s as a surrogate of sociocultural factors underlying item responses is serendipitous in nature. The limitations of investigating the data set in this manner will be discussed extensively in Part 5.2 (pages 276ff.).

In addition to the relationship with demographic variables, another unanswered question from Chapter 3 is whether items of different specs can elicit responses that reflect the same construct (cf. also the rationale, page 186). In the development process, items are generated using target words that are based on a narrow bandwidth of frequency (page 75). In contrast, all but anchor items are constructed in this version anew, using a gradient-frequency test spec (page 188). In light of these differences, the current items will be investigated as to whether they are likely to test the same proficiency as ConCloze 5 items. The procedure leading to placing the items of the two versions in the same test model is called test equating, which can be summarized in Figure 62.

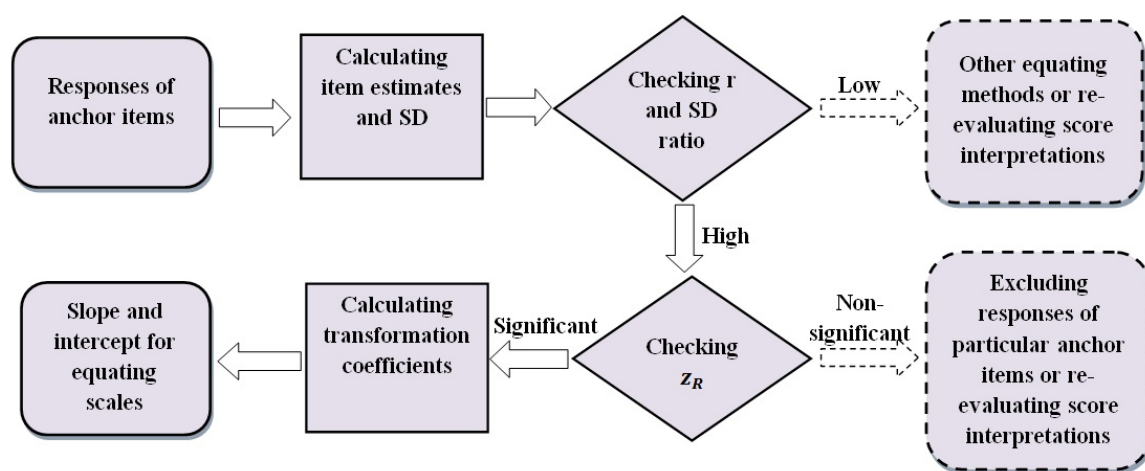


Figure 62 Equating ConCloze 6 to 5

The test equating begins with seeking item estimates (comparable to IFs as used in Chapter 3) and standard deviations of the anchor-item responses in ConCloze 5 and 6. The estimates and standard deviations are then investigated for comparability in variances between the two versions (cf. Linacre 2015b for more details about test equating). This is shown in Figure 62, where r and SD ratio are main statistics for the comparability check between the anchor items as used in the two versions. Item stability (z_R) is also checked

in case particular anchor items should be dropped out of the equating process (cf. Huynh & Meyer 2010: 1 for usefulness of the measure). Afterwards, transformation coefficients are computed, yielding a slope and intercept for modeling a logistic best-fitting equation. Finally, this equation is used for adjusting the estimates of all the ConCloze 6 items, which are plotted alongside those of ConCloze 5 items in Figure 63.

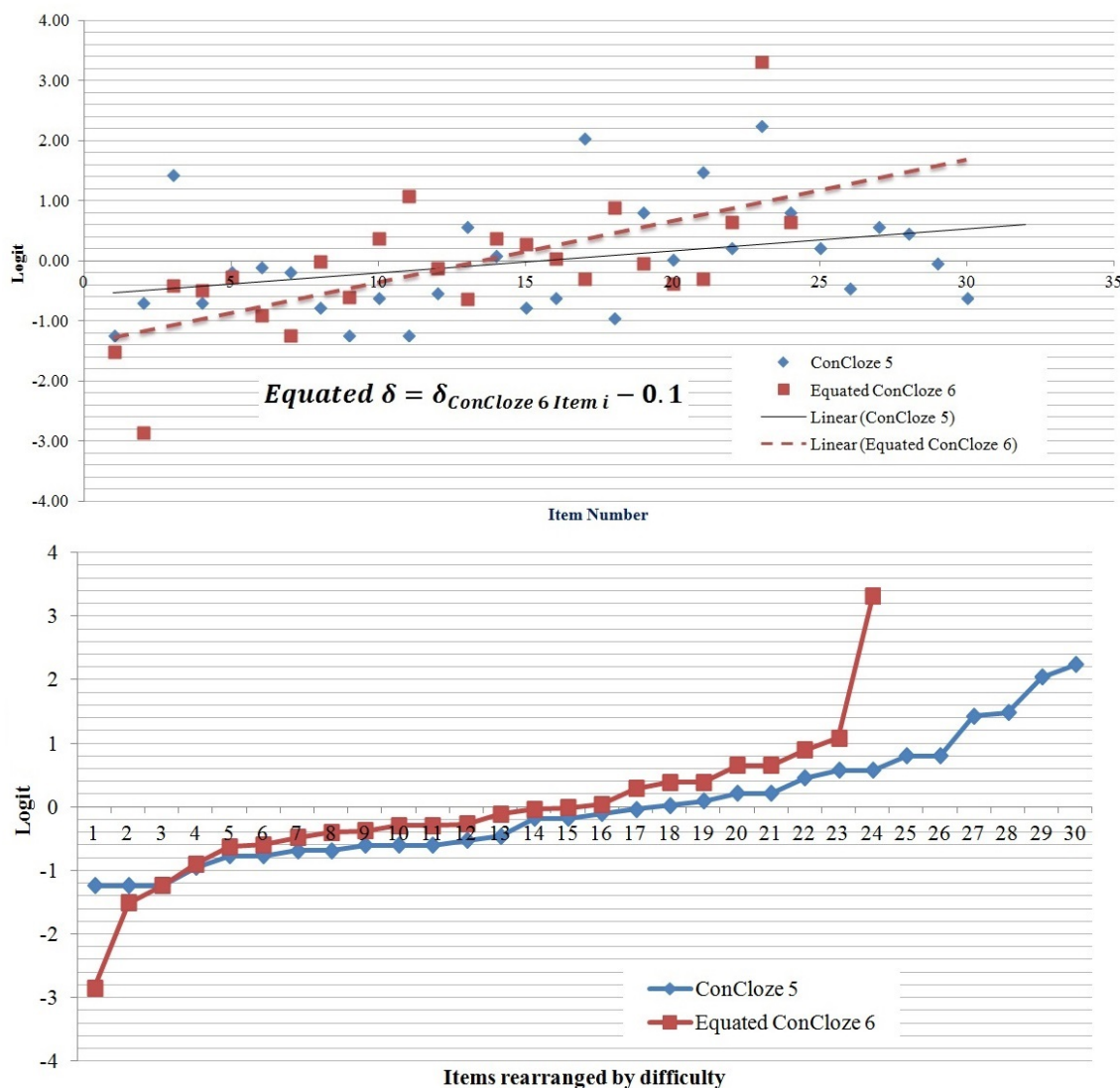


Figure 63 Item estimates of ConCloze 5 and equated ConCloze 6

Figure 63 shows several findings that are potentially significant to the validity investigation. First, the intercept for equating ConCloze 6 to ConCloze 5 is -0.1 . This means that, for equivalent modeling, a -0.1 has been subtracted out of all the item estimates of ConCloze 6. Generally, a maximum deviation of ± 0.3 logit can be deemed

appropriate for assuming that two scales or items are reasonably similar (Linacre 1994). Considering the current intercept, an inference is that ConCloze 5 and 6 are likely to be similar and could potentially measure the same construct proficiency.

ConCloze 5 and 6 are administered separately to different samples of examinees. Yet, a second finding is that their best-fitting lines intersect in Figure 63 above. It can be interpreted that there is a core competence underlying all the item estimates, both in ConCloze 5 and 6. Finding crossing regression lines in the analysis can be important because it is the first log-linear evidence in this study to rely on two differently engineered tests. Usually associated with the term *Rasch modeling*, an advantage is that the evidence can be regarded as relatively independent of the examinee sample used vis-à-vis the sample-dependent measures in Chapter 3 (cf. McNamara 1996: 160ff. for some advantages of Rasch modeling). Considering this, the present evidence increases confidence in arguing that ConCloze 5 and 6 measure the same construct proficiency.

It is worth restating that the convergence that arises between ConCloze 5 and 6 confers a power of generalization of score interpretations. This is so because there being two sets of items means a bigger pool of items sampled and accordingly more observations made on this facet of ConCloze performance. As the finding is based on two separate administrations, it may be deemed interscalar consistency – applicability of score interpretations between scales. On this account, the convergence of the regression lines seems to provide generalizability evidence to the validity investigation.

Test designs which lead to variation in item difficulty indicate differences in processing (page 42). Given this, a third finding is about substantive performance as evidenced by item-difficulty variability, which can be inferred from several differences in observations of Figure 63 above. First, the best-fitting line of ConCloze 6 is steeper than that of ConCloze 5. This indicates that the overall difficulty of ConCloze 6 changes at a faster pace than that in ConCloze 5. Secondly, when the item estimates are rearranged, the majority in ConCloze 6 are also found aligned above those in ConCloze 5. This can be interpreted in the way that ConCloze 6 items are generally more difficult than ConCloze 5 items. Thirdly, it is worth restating that item estimates in ConCloze 6 need a -0.1 subtraction in order to be on a par with those in ConCloze 5. This means that ConCloze 6 items have originally been more difficult than those of ConCloze 5. Finally, the spec of ConCloze 6 features gradient frequency of the options, whereas item

components such as the number of concordance lines are varied by the spec of ConCloze 5. All this evidence indicates that it could be the gradient-frequency design of ConCloze 6 that contributes to a more noticeable change in item difficulty.

Considering the current difficulty variability by item design, the inferences on substantive processing can be as follows. Reducing options, adding more concordance lines, or using semantically unrelated options can lower item difficulty. On the contrary, using options that are semantically unrelated and differ in corpus-based frequency may vary item difficulty relatively efficiently. This means that likelihood of word knownness can be a function of how well the examinees can successfully engage in the test task; the better the examinees know lexical semantics of the options, the more likely they can score.

In light of the convergence between ConCloze 5 and 6, the next investigation is to calibrate their items together. Calibrating ConCloze 5–6 items together despite their different item specs can be useful because the calibration could produce additional types of measure that may reveal underlying patterns of task processing. Moreover, all items, rather than anchor items only, can be modeled for substantive association between the two scales. For example, a group of items may systematically elicit different response patterns, suggesting a dissimilar competence at work. Accordingly, there being more items in each group of test items and eventually in one test modeling means that such a difference in response patterning could be identified easier. To that extent, anomalies in item responses are also more likely to be detected than in investigating a scale of a smaller number of items.

Considering the interscalar consistency discussed earlier, an equivalent-groups design (Dorans *et al.* 2010: 9) and concurrent-equating method (Linacre 2015a) are followed in co-calibrating ConCloze 5–6 items. Table 40 below shows the calibration result, in which the responses to the four anchor items from ConCloze 5 and 6 are merged. The log-linear estimates of 78% of the items (39 of 50) are found to lie between -1 and 1 . This indicates that the majority of them are moderately difficult. Accordingly, an inference based on this calibration is that the difficulty level of the task content is likely to be appropriate for the population.

Table 40 ConCloze 5–6 concurrent item measures and fit statistics

Item	Source of Item Responses*	Measure	Std. Error	Weighted Mean-square (WMS)	Std. WMS	Unweighted Mean-square (UMS)	Std. UMS
1	ConCloze 5 Item 1	-1.27	0.32	0.99	0.03	0.88	-0.13
2	CC5 I2	-0.73	0.28	0.82	-1.17	0.69	-0.93
3	CC5 I3	1.39	0.24	1.24	2.4‡	1.5	2.5‡
4	CC5 I4	-0.73	0.28	0.78	-1.44	0.65	-1.11
5	CC5 I5	-0.21	0.26	0.88	-0.96	0.79	-0.84
6	CC5 I6–CC6 I9	-0.45	0.14	0.85	-2.2	0.74	-1.8
7	CC5 I7	-0.21	0.26	0.99	-0.07	0.85	-0.54
8	CC5 I8	-0.81	0.29	1.13	0.79	1.19	0.65
9	CC5 I9–CC6 I6	-0.98	0.16	0.81	-2.39	0.58	-2.34
10	CC5 I10	-0.65	0.28	1.02	0.2	1.26	0.88
11	CC5 I11	-1.27	0.32	0.89	-0.51	0.82	-0.27
12	CC5 I12	-0.57	0.28	0.89	-0.69	0.74	-0.84
13	CC5 I13	0.53	0.24	1.27	2.5‡	1.36	1.93
14	CC5 I14	0.05	0.25	0.95	-0.37	0.93	-0.27
15	CC5 I15	-0.81	0.29	0.96	-0.19	0.8	-0.51
16	CC5 I16	-0.65	0.28	1.08	0.53	0.99	0.06
17	CC5 I17	2	0.26	1.22	1.83	1.64†	2.37‡
18	CC5 I18–CC6 I7	-1.15	0.16	0.96	-0.39	1.02	0.17
19	CC5 I19	0.76	0.24	1.01	0.12	1.01	0.13
20	CC5 I20	-0.01	0.25	1.21	1.72	1.39	1.66
21	CC5 I21	1.45	0.24	0.95	-0.54	0.91	-0.44
22	CC5 I22	0.17	0.25	0.9	-0.88	0.8	-1.02
23	CC5 I23	2.2	0.26	1.23	1.72	1.15	0.64
24	CC5 I24	0.76	0.24	1.01	0.13	1.01	0.13
25	CC5 I25	0.17	0.25	0.91	-0.75	0.82	-0.86
26	CC5 I26–CC6 I8	-0.12	0.14	1.02	0.35	1.12	0.91
27	CC5 I27	0.53	0.24	1.3	2.72‡	1.36	1.93
28	CC5 I28	0.42	0.24	0.75	-2.61	0.65	-2.19
29	CC5 I29	-0.08	0.26	0.84	-1.31	0.8	-0.85
30	CC5 I30	-0.65	0.28	0.92	-0.46	0.8	-0.59
31	CC6 I1	-1.49	0.2	1.03	0.28	0.93	-0.1
32	CC6 I2	-2.84	0.32	0.86	-0.47	0.38	-1.01
33	CC6 I3	-0.38	0.17	0.91	-1.21	0.8	-1.13
34	CC6 I4	-0.46	0.17	1.08	1.07	1.24	1.22
35	CC6 I5	-0.24	0.16	1.14	1.78	1.11	0.68
36	CC6 I10	0.41	0.16	1.1	1.39	1.04	0.4
37	CC6 I11	1.1	0.16	1.1	1.42	1.16	1.45
38	CC6 I12	-0.09	0.16	1.07	0.93	1.31	1.87
39	CC6 I13	-0.6	0.17	0.9	-1.18	0.75	-1.26
40	CC6 I14	0.41	0.16	1.13	1.91	1.12	0.97
41	CC6 I15	0.31	0.16	0.86	-2.11	0.8	-1.66
42	CC6 I16	0.06	0.16	0.87	-1.92	0.76	-1.82
43	CC6 I17	-0.27	0.16	0.96	-0.49	0.93	-0.34
44	CC6 I18	0.91	0.15	1.15	2.11‡	1.16	1.46
45	CC6 I19	-0.01	0.16	0.87	-1.86	0.72	-2.08
46	CC6 I20	-0.35	0.16	0.93	-0.91	0.79	-1.2
47	CC6 I21	-0.27	0.16	0.91	-1.28	0.75	-1.59
48	CC6 I22	0.67	0.15	1.28	3.82‡	1.5	3.94‡
49	CC6 I23	3.34	0.22	1.18	1.3	2.26†	3‡
50	CC6 I24	0.67	0.15	0.95	-0.66	0.88	-1.12
\bar{x}		-0.0008					

* Item responses of the anchor items are merged. The item responses of the rest are modeled with the missing responses treated as ‘ignored’.

† indicates a mean-square value that are unproductive to the test model (Linacre 2002). The present criterion is > 1.5 .

‡ indicates a standardized mean-square value that is noticeably unpredictable by the test model (ibid.). The criterion is > 2 .

In light of the inference about task difficulty, response invalidity that may arise from an inappropriate difficulty level seems to be improbable. Moreover, no pattern that is specific to either ConCloze 5 or 6 can be observed from the item estimates and fit statistics (detailed later). Nor can any systematic anomaly be detected in the statistics of either anchor items or the rest of the items. This means that almost all of the items could be aligned on the same construct model. Finding the unanimity in model alignment has two implications. First, it indicates consistency of two different sets of items which are administered separately. This offers a piece of generalizability evidence for the validity investigation. Secondly, the finding also accords with the unidimensionality discovered in the test equating (cf. page 206). This means that the construct proficiency can be observed consistently in two different facets of construct measurement (analyses of individual scales and concurrent calibration), thereby giving increased confidence in the test-score interpretations.

Apart from the information about moderate item difficulty and scale unidimensionality, other information in Table 40 above is on misfit items. In Rasch modeling, items are expected to be definable along a unidimensional construct (Linacre 2004). When item responses are noticeably deviant from the construct model, the items are considered misfit. Learning features from a misfit item could be useful because it may provide discriminant-validity evidence for construct definition. The misfit items in Table 40 can be categorized into three groups, which will be discussed in turn.

The first group of misfit items comprises those that are underfit only in terms of standardized weighted mean-squares (Std. WMS). These include ConCloze 5 Items 13 and 27, and ConCloze 6 Item 18. For brevity, the responses in only one item, Item 13, are illustrated in Figure 64.

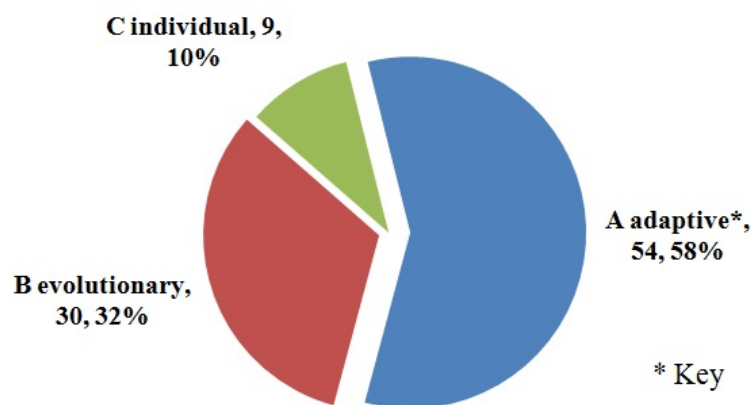


Figure 64 ConCloze 5 Item 13 responses

In ConCloze 1, some distractors whose semantic elements are closely related with those of the target words are found receiving a competing number of responses (page 98). An interpretation then is that lexical-semantic knowledge is likely to be tapped into during task engagement. In Figure 64 above, the distractor *evolutionary* receives a large number of responses, approximately half the size of that of the target word *adaptive*. *Evolutionary* refers to an ability to change over time, whereas *adaptive* also denotes an ability to change. This semantic relation is close and may help to explain the attractiveness of the distractor. An explanation is that the semantic closeness may potentially make this distractor difficult to distinguish from the target word, so much so that it draws some responses from the examinees in the higher-ability group, causing unmodeled noise and accordingly an underfit Std. WMS. For substantive processing, this means that the ConCloze examinees actively interpret lexical semantics of the options. Testing semantic compatibility of a given word in the concordance context is likely to be a definite part of the competence required (pages 136f.). Yet, differentiating close semantic components could be peripheral to the construct proficiency and would pose some difficulty in engaging with the test task.

The second group of misfit items comprises those items that are underfit in Std. WMS and standardized unweighted mean-squares (Std. UMS). These include ConCloze 5 Item 3, and ConCloze 6 Item 22. For brevity, the responses are illustrated for only one item, Item 3, in Figure 65 below. It is found that the distractor *seek* receives a large number of responses, nearly the same size as that of the target word *elicit*. *Seek* can refer to trying to obtain something, whereas *elicit* also denotes obtaining something. This semantic relationship is close and may help to explain the attractiveness of the distractor

seek. An explanation is that, again, the semantic closeness may potentially make this distractor difficult to distinguish from the target word. This difficulty could be to the extent that it draws some responses from the examinees in the higher-ability group—causing noise in measurement and accordingly an underfit Std. WMS—while some examinees in the lower-ability group may guess correctly, causing an underfit Std. UMS. For substantive processing, this means that, again, the semantic relation among the options could be processed by ConCloze examinees. When two options are close in meaning, difficulty in task processing likely arises for both higher-ability and lower-ability examinees.

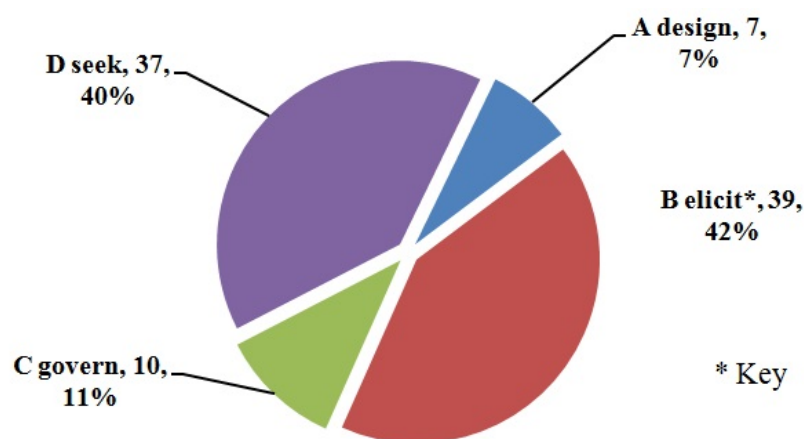


Figure 65 ConCloze 5 Item 3 responses

The last group of misfit items in Table 40 (page 209) belongs to those that are underfit in UMS and Std. UMS. These include ConCloze 5 Item 17, and ConCloze 6 Item 23. For brevity, the responses to only one item, Item 23, are illustrated in Figure 66 below. This item is selected because its UMS is 2.26, which can be considered abnormally large (cf. Wright & Linacre 1994; Linacre 2002 for detailed interpretations). Its item estimate is also very high at 3.34, suggesting very high difficulty. Besides, its Std. UMS is also grossly underfit at 3. An item which is very difficult and whose Std. UMS is underfit can be considered a bad item (Linacre ca. 2012). This is in the sense that the responses are too unpredictable for the test model. An interpretation is that some item responses of Item 23 are grossly distorting to the test model.



Figure 66 ConCloze 6 Item 23 item responses

Ambiguous or misleading options can cause underfit mean-squares (Linacre ca. 2012). Given this, a hypothesis is that misleading semantics of option words could be a source of the distortion in Item 23. In Figure 66 above, the distractors *mitigating* and *paternalistic* attract 30% and 44% of the item responses, respectively. By contrast, the target word *piecemeal* draws only 16%. This is a marked contrast in response pattern between the target word and distractors, in which the target word fails to amass the majority of responses. Considering the failure of the target word, its meaning aspect could be ambiguous and cause construct-irrelevant variance: the target word *piecemeal* denotes something that ‘happens slowly and in stages that are not regular or planned properly’ (Longman Dictionary of Contemporary English 2015). This meaning does not seem to be composite—viz. not *piece* + *meal*—and may involve a false-friend process similar to “morphological troublemakers” in lexical guessing (Bensoussan & Laufer 1984: 29).

In light of the finding of potential false-friend turbulence, the implications for this study are twofold. A first implication is for substantive processing. In ConCloze 5, semantic relationship among the option words is found to have a powerful effect on item difficulty (page 170). That a non-composite meaning of an option word could be a source of turbulence in the current construct model seems to confirm that finding. It indicates that the respondents really contrast lexical components of the options actively during task engagement. An inference is that when one of the options, especially the target word itself, has a possibly unknown or unguessable meaning, the usual semantic contrast could be toppled.

The second implication of the false-friend turbulence is for domain description. The lexical-semantic relation among option words, again, has been found to be critical in task engagement (page 170). Up to this stage of the research, the relationship can work out in three possibilities. First, when semantic components of option words are known to the respondents, recognizing word information embedded in the concordance prompt as associated with the target word is likely to lead to a successful task completion. This is when proficient respondents can contrast those components considering the lexical network. They could hence have a higher success rate across items in the entire test than the respondents with low proficiency levels. Second, when a pair of option words share some of their semantic components, a challenge arises in using the word information for deciding between them (cf. also their moderating effects in ConCloze 5, page 171). This is when the test task could retain its difficulty even though more concordance lines are added for item modification, for instance. Lastly, when semantic components of an option word misleads the respondents, they—as evidenced by the current fit turbulence—may compensate for this misunderstanding by choosing other options. This implies that the knowledge of word associations induced by the context clues can be one-directional: seeing the clues enables the respondents to associate with particular choices, but not vice versa when the choice is unknown or misunderstood.

In addition to analyzing the fit statistics, exploring the item estimates by the Rasch modeling in Table 40 (page 209) can also address a primary question on effects of corpus-based frequency on item difficulty. In ConCloze 1, the frequency level of the target words is not found to singly drive item difficulty (page 95). The inference then is that examinees must have processed other item components apart from the target words. In Figure 67 below, the item estimates of ConCloze 6 from Table 40 are rearranged to the frequency bands. A mixed impression is found to arise from the pattern of option-frequency effects. On the one hand, the 0–0.3K frequency band is lower than and is almost completely separated from the 2.4–2.7K band. This allows an interpretation that the former items are nearly totally easier than the latter ones. On the other hand, three item estimates of the 1.5–1.9K band are lower than their counterparts of the 0.7–1.1K band. An interpretation is that the majority of the lower-frequency items are easier than the latter ones. It can thus be inferred that options of a lower frequency do not necessarily increase item estimates, and the contrary. In light of this evidence, it may be argued that the frequency level of the options has only a moderate effect on item difficulty.

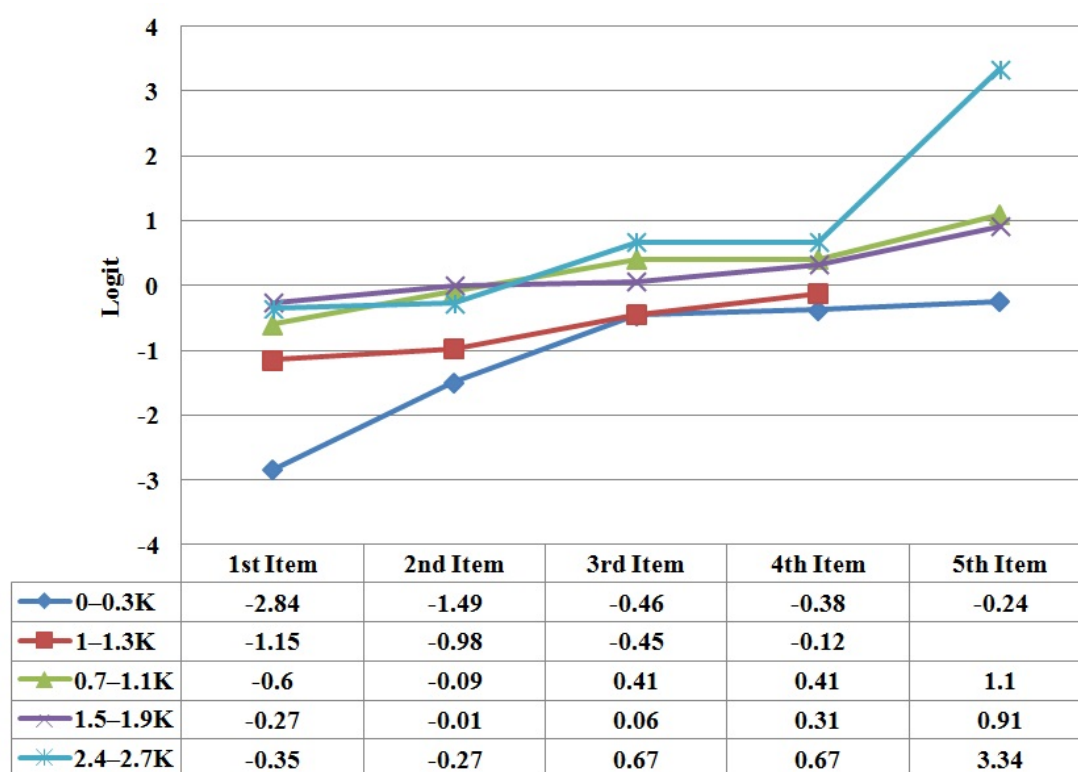


Figure 67 ConCloze 6 frequency levels and ascending item estimates

Further, the ways the frequency lines in Figure 67 are aligned also reflect the moderate effects of word frequency. First of all, the spec is designed with interspersed exclusions in between the frequency bands (cf. pages 188f.). However, the lines of the three upper bands, viz. 0.7–1.1K, 1.5–1.9K, and 2.4–1.1K, seem clustered. This means that their item estimates do not differ decisively, thereby confirming the inference that the frequency level of options is not the only determining difficulty driver. Accordingly, the interim hypothesis regarding the influence of corpus-based frequency over item difficulty (page 186) is accepted with reserve.

In light of the non-dominating effects of corpus-based frequency, the co-influentiality of multiple test-task components during engagement is also confirmed. A substantive summary could be as follows. ConCloze examinees generally process the semantic components of option words concomitantly. In turn, the options are processed in relation to the context clues available in the concordance prompt. In less proficient examinees, some or all of the option words are unlikely to be fully comprehended. They

would be unable to decipher the very meanings of these option words, let alone their semantic components and word associations that may be found in the concordance prompt. Therefore, it could be argued that the interpretability of option-word meaning could be a joint function of difficulty with the corpus-based frequency.

A high corpus-based frequency is generally indicative of the likelihood of word knownness (cf. an earlier argument about a non-linear mental lexicon on page 96). Given the co-influentiality of multiple task components processed, the nature of the knowledge tested in ConCloze can also be inferred. First of all, words of the options are unlikely to be stored in the mental lexicon in accordance with the observed corpus-based frequency. As discourse structure and content generally vary from text to text, their meaning would not be retrieved accordingly from the mental lexicon. An implication is that the lexicon is activated selectively by the word associations that appear in a text, which in this study could be the concordance prompt.

4.2.4 WAF–ConCloze relationship

A research hypothesis is that ConCloze scores will correlate significantly with WAF scores (page 16ff.). This is based on two primary reasons. First, in the theoretical framework for construct measurement, WAF and ConCloze are theorized to likely have several aspects of substantive content in common (cf. pages 47ff.). Albeit unlikely to test exactly the same domain of competence, having some substantive content in common suggests a potentially positive result in correlation analysis. A second reason is test logistics—such as availability, and simplicity in administration and delivery—indicating practicality in using the test as a criterion test (*ibid.*). Figure 68 below shows a scatterplot of the total scores by the examinees who finished the test battery. The correlation coefficient between the scores is significantly positive ($r(237) = 0.78, p = .000$). The r squared is 0.61, indicating high co-variability. A positive regression line can also be drawn through the plots of the scores. These pieces of evidence are important because they are among the first criterion-related evidence in this study.

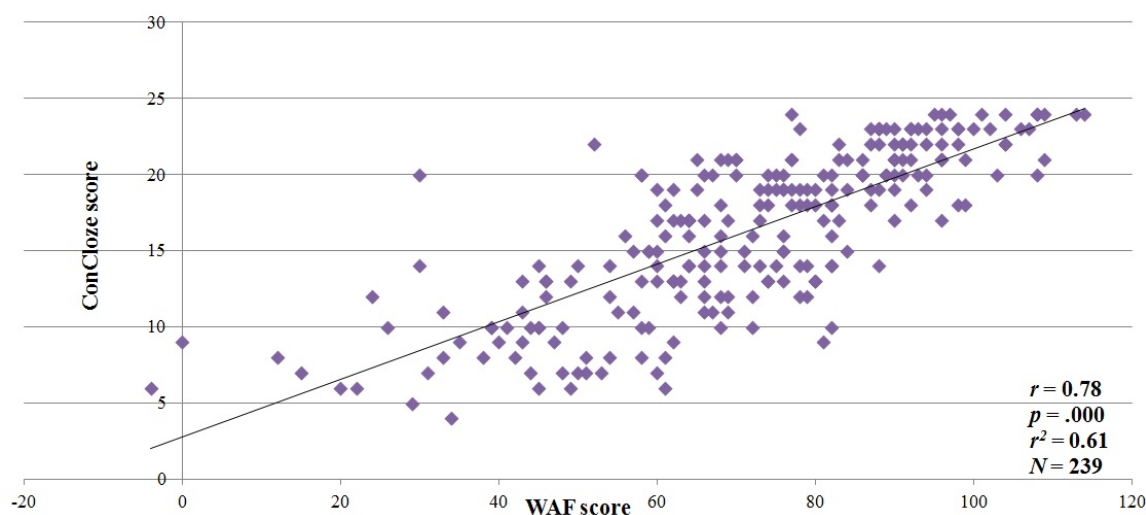


Figure 68 Scatterplot of ConCloze 6 and WAF total scores

Given that positive criterion-related evidence is obtained, three key construct-related inferences may be made. First, when the examinees perform well in ConCloze, it is likely that they also perform well on WAF, and vice versa. A second inference is about the competence domains tested in ConCloze. WAF is known to test knowledge of word association (cf. Figure 13 (page 48) for an item example), which includes knowledge of synonymy—lexical-semantic aspects such as meronyms and holonyms included—and knowledge of collocation (Read 1993; Schmitt 2010: 226ff.; Schmitt *et al.* 2011). In light of the positive relationship, an exploratory inference can be that the ConCloze construct is likely to encompass the knowledge of word association. On the one hand, the ConCloze construct is hypothesized to be context-based multicomponentiality (pages 35f.). Yet, finding the possibility of one domain subordinating two more language-related subdomains is an encouraging sign of multicomponentiality. On this account, the hypothesis mentioned earlier (page 216) can be accepted. The final key inference is about general proficiency in English. WAF is reviewed in Chapter 2 (pages 47f.) for its positive relationship with general proficiency in English. In light of the positive correlation, a likelihood is thus that performance in ConCloze would vary positively with general proficiency in English. An implication is that ConCloze scores may usefully reflect the level of general English proficiency as well.

In addition to analyzing the correlation, factors underlying the item responses will also be investigated. Exploring underlying factors between ConCloze and WAF can be useful for three reasons. First, in the previous Rasch modeling, for example, item

variances are tested against a preconceived unidimensional model (pages 205ff.). By contrast, analyzing factors can let the variances speak for themselves, modeling the principal components in a hypothesis-free fashion (cf. also Hupé 2015 for some caveats of permutating data for latent factors). This means that it can extract a minimum number of factors, rather than presume a single construct, which may maximally account for the response variances. ConCloze and WAF could also be found to share a single factor, some factors, or no common factor, which would reflect differing dimensions of their association in detail. It is worth stating that the terms *latent factor*, *principal component*, and *root* will be used interchangeably due to their common reference in the current analysis.

A second reason in support of analyzing underlying factors is to estimate random errors. The factor analysis that is going to be performed incorporates a large-scale simulated permutation of item responses; seemingly subtle variations in response patterning such as outliers will be amplified through the simulation. This means that if the variations are in fact construct-related, their patterns would become discernible. The amplification also applies to the latent factor(s) of ConCloze associated with WAF, if any.

The last reason for the present analysis is to fine-tune the ConCloze–WAF relationship. In WAF, knowledge of synonymy and knowledge of collocation are tested (pages 47f.). However, the previous correlation analysis (page 217) cannot determine if the responses in both of these categories correlate equally with the responses to ConCloze. To that extent, it cannot be specified yet if it is only the collocates, synonyms, or both that govern the strong correlation. Investigating latent factors can thus provide fine-tuned interpretations complementing the correlation finding.

The process of analysis can be summarized in Figure 69 below. It starts with arranging all the ConCloze and WAF items separately according to their difficulty. This is equal to their item estimates, as in item measures of Table 40 (page 209). The items are then clustered to form item parcels (IPs). The aim is to prepare items of similar difficulty levels, the responses of which will then be analyzed together. Also shown in Figure 69, four IPs are formed for ConCloze, containing five items each, with the anchor items excluded. Likewise, WAF IPs are formed based on the item estimates. This begins with treating WAF items as super-items and the key options (i.e., the correct answers, e.g.,

helpful, positive, response, and weather in Figure 13 (page 48)) as subitems. Yet, a crucial distinction is that the distractors in WAF (e.g., *habit, legal, possible, and teacher* in Figure 13) are excluded altogether because of their substantive processing that may not be comparable to that in ConCloze. WAF is good for indicating what the examinees know rather than what they do not know (Read 2012: 310f.). This means that the keys in WAF involve selecting or not selecting themselves, and the distractors are not rejected actively. This substantive performance is similar to that in ConCloze, (a) which entails selecting or not selecting the key, and (b) in which each response may not always reflect the ability to reject distractors knowingly. The process of parceling WAF subitems produces 13 IPs for the collocate category, and 11 IPs for the synonym category, as summarized in Figure 69.

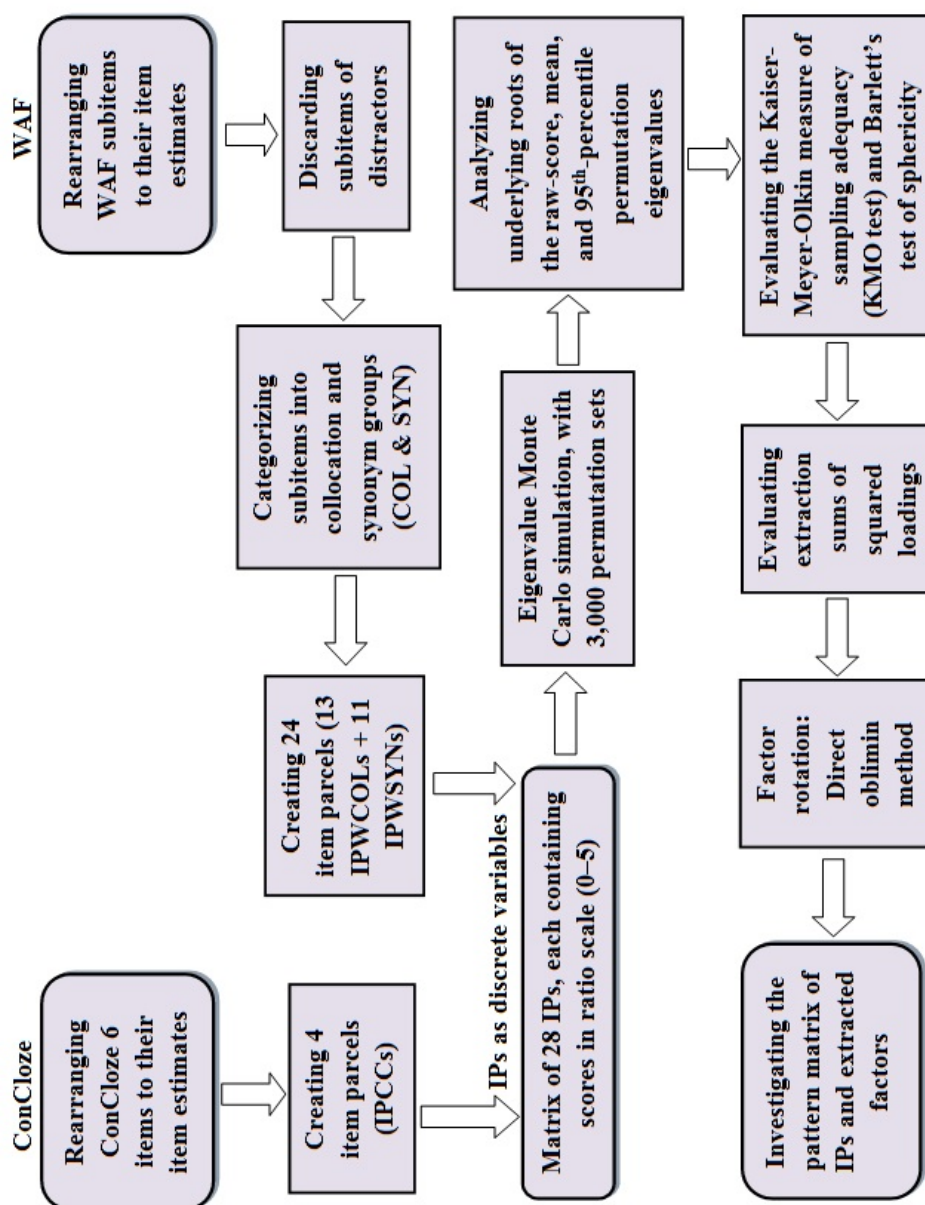


Figure 69 Analyzing principal components of ConCloze-WAF item responses

Figure 69 above also shows the grouping of ConCloze IPs and WAF IPs. Each IP serves as a discrete variable, the responses in which have been added up as the score (cf. Jasper 2010; Batty 2012 for details of this technique). For example, one WAF IP contains five responses to five subitems. If three of these responses are correct (i.e., selecting the keys), the score for this IP is three. The process thus far is crucial because it changes polytomous responses (–4 to 4) of each WAF super-item into a discrete score (0 to 5) in the ratio scale. The same is true with each ConCloze IP, in which dichotomous responses (0 or 1) are merged to form a discrete score (0 to 5). Then, as displayed in Figure 69, the

scores of all IPs are simulated in an eigenvalue Monte Carlo parallel analysis (Buja & Eyuboglu 1992; Franklin *et al.* 1995). The idea is to permute the IP scores randomly such that the eigenvalues of the random-permutation matrices (here mean and 95th-percentile) can be compared with those underlying the actual scores. The result of 3,000 permutation sets is shown in Figure 70 below, in which four roots of the observed scores have greater eigenvalues than those of the mean and 95th-percentile score permutations. An interpretation is that four is a prospective number of latent factors accounting for the majority of the score variances of the entire battery.

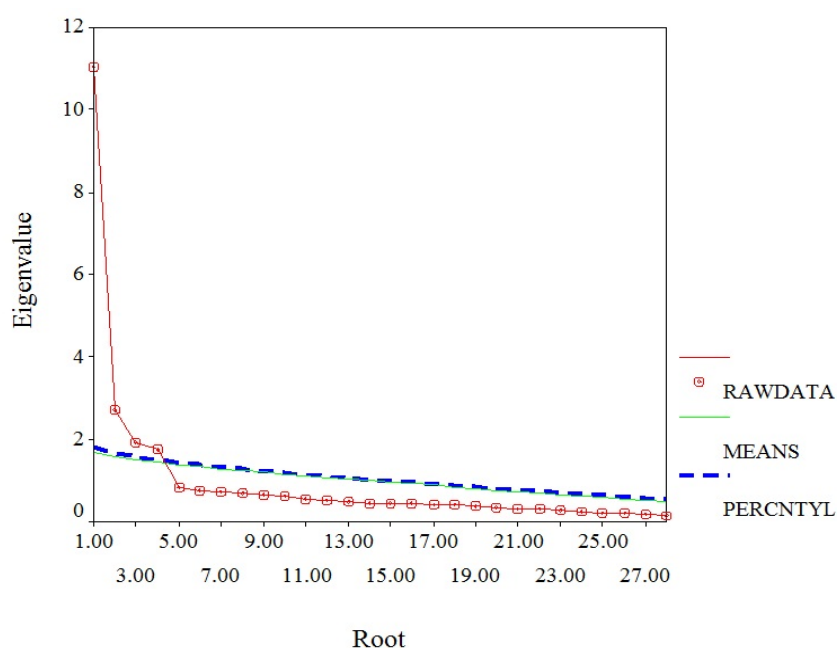


Figure 70 ConCloze 6 and WAF eigenvalue Monte Carlo simulation

The next step in Figure 69 is to check factorability of the IP correlation matrix. In summary, the correlation matrix is compact and not collinear; up to 62% of the regression coefficients can load on four principal components. The process culminates in Table 41 below, which shows a pattern matrix of the IPs and the components derived. The variances of all the ConCloze IPs are found to load heavily on Component 2. The loading is not shared significantly with any of the WAF IPs. Nor does any ConCloze IP load heavily on specific components of the WAF IPs. An interpretation is that given one unit of change in any of the WAF components, ConCloze scores do not vary much. Despite ConCloze being designed for comparability with WAF (page 192), the two tests do not

seem to share any identical underlying factor. Accordingly, the hypothesis that their relationship is significantly positive is rejected.

Table 41 Pattern matrix of IPs and extracted PCs

IP	Component* (Underlying Explanatory Factor)			
	1	2	3	4
IPWCOL16	.816	.007	-.066	-.047
IPWCOL17	.738	.057	-.098	-.111
IPWSYN28	.727	.023	.118	.097
IPWCOL13	.680	.031	-.238	-.097
IPWCOL15	.645	.106	-.175	-.010
IPWCOL14	.610	-.078	-.307	-.013
IPWSYN25	.495	-.064	-.065	.385
IPWSYN23	.491	-.056	.112	.420
IPWSYN27	.467	.011	-.066	.284
IPWCOL12	.463	.078	-.363	.062
IPWSYN26	.392	.089	-.130	.282
IPCC3	-.023	.829	-.002	.029
IPCC2	.003	.806	-.018	.007
IPCC1	.028	.805	.096	.083
IPCC4	.074	.790	.031	-.092
IPWCOL5	-.089	-.004	-.933	.049
IPWCOL7	.070	-.014	-.841	.019
IPWCOL6	.024	-.035	-.831	-.035
IPWCOL9	.113	-.001	-.791	.023
IPWCOL8	.077	-.126	-.733	.176
IPWCOL10	.371	.039	-.588	-.029
IPWCOL11	.287	.092	-.566	.127
IPWSYN19	-.094	.052	-.073	.794
IPWSYN18	-.245	.083	-.280	.726
IPWSYN22	.210	.022	.127	.661
IPWSYN24	.353	-.120	.009	.560
IPWSYN20	-.067	.187	-.369	.551
IPWSYN21	.296	-.122	-.125	.527

* Rotation method: direct oblimin with Kaiser normalization; rotation converged in 11 iterations

That the hypothesis about shared underlying factors between ConCloze and WAF is rejected carries three implications. First of all, interpreting a correlation coefficient has subtle variations. It can denote either a common element (such as Greek and Latin scores) or a common cause (Kline 1991: 5). Given this, a first implication is that the high correlation coefficient previously discussed (pages 217f.) and the present lack of co-factoriality may not be in conflict. Rather, they indicate two layers of governing competence, the superordinate of which could be a general factor in human intelligence. On the one hand, the notion of an overarching general-ability factor such as *g* is as yet

invalid (Vollmer 1983; Morgan 1992; cf. also Oller 1979: 423ff. for a proponent argument for the notion). Yet, the idea of a general verbal ability is not unheard of (e.g., Steele & Aronson 1995: 799). Hence, general proficiency in vocabulary use can be the common domain to which the positive correlation is attributed.

In addition to the superordinate domain, a second implication is about the scale-level constructs of ConCloze and WAF. In the framework for construct measurement (Section 2.3, pages 48ff.), the two tests are predicted to have the domains of collocation and synonymy in common. However, no empirical co-factoriality is discovered here in either respect. For example, if ConCloze examinees processed the network of collocates in the concordance rigorously, then the ConCloze IPs should have been found to load on the same components as the collocate IPs of WAF. This implies that there is more in the lexical network of the concordance prompt for the examinees to process than collocates and synonyms.

Another related implication is about discriminant evidence. Vocabulary use involves many different aspects of using words properly, e.g., meaning, collocation, referents, association (cf. also Table 3, page 34 for a widely-cited classification of aspects of knowing a word receptively). Given this, the scale-level constructs of ConCloze and WAF could differ in the main aspects involved; the ConCloze item type could test a unique combination of aspects of proficiency in vocabulary use. It is also worth considering that the ConCloze IPs load heavily on a single component—a sign of internal consistency providing structural-validity evidence—suggesting a model of a single latent factor (cf. Messick 1989; John & Benet-Martínez 2000). Accordingly, the present evidence of factor loadings discriminates the ConCloze construct from that of WAF by (a) not loading on the WAF factors, and (b) exclusively loading on their own factor.

In light of the uniqueness, a third implication lies with the substantive processing of ConCloze. It is inferred earlier (page 217) that the construct is likely to include knowledge of word association. Then the factor loadings above indicate there could be more for the ConCloze examinees to process than collocates and synonyms. Accordingly, an inference is that the underlying processes between the two tests are inherently different, to the extent that the knowledge of word association mobilized in the ConCloze tasks would expand beyond those relating to collocation and synonymy. Moreover, it is worth considering that WAF examinees are asked to choose up to four options during

task engagement (cf. Figure 13 (page 48) for an item example). They are not required to choose one word over the others as long as their selections are not over four. By contrast, ConCloze examinees are asked to choose only one option over the others. The previous findings also indicate that they process the very relations existing among the option words, rather than merely the semantic components of each option word irrespective of those of the others (e.g., pages 97f.). On these accounts, actively weighing up one option over the others could be a decisive distinction in task engagement between the two tests. The substantive performance in ConCloze could thus be summarized as follows. At least knowledge of synonymy, collocation, and lexical-semantics operate, in which the semantic relations among the option words are critical in testing their compatibility with the lexical network of association formed by the concordance-based clues. It could also be argued that the score meaning is composite.

4.2.5 Inferences and decision

In Chapter 3 (page 180), the research question on whether ConCloze scores are associated with WAF scores is evaluated as unanswered. In this part, two pieces of evidence are collected to address the question. One is a correlation coefficient between the two tests, which shows a positive association between them. The other is an analysis of underlying components, in which the two do not have any in common. Major evidence and inferences are summarized in Table 42 in no particular order.

Table 42 Major ConCloze 6 evidence and inferences

Entry	Evidence	Inference
1	Improved completion rate	Low response invalidity
2	Negative-skewed score distribution in both the battery and individual-scale levels	Appropriate difficulty level, low response invalidity
3	Repeated pattern of stronger association of the scores with education level than with age	Time-consuming acquisition of the competence in formal education
4	Repeated pattern of stronger association with English self-ratings than with education level and age	Proficiency being fundamental to English skills
5	Repeated pattern of strong association between self-ratings and scores	Lexical-semantic knowledge as part of the construct, positively associated with general English proficiency
6	Repeated pattern of strongest and second strongest associations with the reading and writing self-ratings, respectively	Construct of lexical-semantic knowledge in English reading and writing
7	Systematic regression residuals	Much of the score variance explained by neither of the demographic variables
8	L1 variable with a very large η_p^2	Construct proficiency dependent on the extent and intensity of exposure to English
9	Intercept of -0.1 intercept and crossing regression lines in test equating	Interscalar consistency and generalizability of score interpretations
10	ConCloze 6 items generally more difficulty than ConCloze 5 ones	Scale difficulty varied more efficiently by amplifying option-frequency levels than by varying IV features (e.g., increasing concordance lines)
11	Concurrent calibration without systematic anomaly by item design	Appropriate item difficulty for the intended population; structural validity for unidimensionality interpretation
12	Misfit items with option words of related semantic components and of morphological trouble-making	Examinees actively interpreting the lexical-semantics (semantic components) of option words
13	Indecisive effects of option-frequency levels on item difficulty	Co-influentiality of multiple test-task components and selective pickup of word associations in the prompt
14	Significantly positive correlation coefficient between ConCloze and WAF	Knowledge of word association encompassed in the construct
15	ConCloze IPs loading on a different principal component from WAF IPs	ConCloze testing a unique combination of aspects in proficiency in vocabulary use

In addition to dealing with the remaining issue about criterion-related validity, evidence on association with demographic variables is also examined in this part. This is an investigation of inadequacy which only one administration in Chapter 3 (ConCloze 5) has provided evidence for. Almost all of the correlation patterns are found similar to the counterparts in ConCloze 5 and accordingly suggest adequacy in this respect for the validity investigation. This means that adequate evidence is currently accumulated for nearly all the investigations in Table 32 (page 180). Appraising the entire validity

investigation, it seems that the score interpretations obtained through the item type have validity.

In Chapter 1, the test purpose is hypothetically set to be a proficiency test on professional and academic English grammatical and vocabulary use (page 23). In light of the summary, the test purpose can be appraised as follows. Up until ConCloze 5, the test purpose has been narrowed down to a proficiency test on vocabulary use, with emphasis on lexical-semantic knowledge, and knowledge of word association. With the findings of this version added up, the ConCloze item type seems to test a unique combination of knowledge domains of proficiency in vocabulary use. Specifically, while knowledge of collocation and knowledge of synonymy could be subordinate to the knowledge of word association tested in the test format, the findings seem to indicate that there is more in the word information that is actively processed and hence tested in the item type. In sum, the construct domains tested in ConCloze are encompassed in proficiency in vocabulary use.

Notwithstanding adequacy of the evidence appraised earlier in Table 32 (page 180), one investigation in the test-development process still remains marked as inadequate; substantive processing has thus far failed to obtain another piece of evidence in this ConCloze version. Considering this, the validity inquiry should continue in this respect. In doing so, it is decided that variation should also be added to the test method because the test versions from ConCloze 1–6 are all based on the multiple-choice format. Finding congruence between test methods could give generalizability-validity evidence, thereby increasing confidence in test interpretations.

4.3 ConCloze 7: Substantive Fine-tuning

4.3.1 Rationale

A decision from ConCloze 6 (page 226) is to continue this study with investigating substantive processing. Examining the processes underlying test-task engagement can be useful for two reasons. First, evidence in this respect has been evaluated as inadequate; the investigation can hence address this inadequacy. Secondly, the score interpretations made thus far can be fine-tuned from the qualitative viewpoint. Representing multi-faceted observations, such a fine-tuned interpretation would confer a

power of generalization and add further utility to the construct inferences (cf. also Figure 2 (page 2) for the model of test validity).

Generalizability of validity evidence is when a construct-related inference is applicable across occasions (cf. also the research question on the generalizability aspect of construct validity on pages 16f.). Figure 71 below shows the current inquiries into facets of construct measurement for the generalizability aspect. For example, a third inquiry is to devise another item format called *modified constructed-response*. It elicits open-ended responses, which are another type of response expected of the examinees. The underlying processes and strategies of this format will be investigated as to whether they are comparable to those of ConCloze 1–6. Comparability can serve as evidence representing another facet of test method, suggesting generalizability of the test interpretations.

<i>Generalizability-validity Evidence of the ConCloze Score Meaning</i>		
Other task content: Verbalizations of different items	Another analytical approach: Face validity incorporated	Another type of expected response: Modified constructed-response format

Figure 71 Investigating substantive processing for a generalizability argument

Each inquiry in Figure 71 involves a distinct set of test items. As such, ConCloze 7 refers not to one item set but to a collection of three inquiries into different facets of test method. For this reason, ConCloze 7 does not have a separate section of the test spec and hence is also structured so as to reflect the multi-method nature. For brevity, an iterative presentation is used, in which the rationale for each will be first provided. Then the inquiry process will be summarized, followed by the results of the inquiry. At the end of this part (4.3), concluding claims will then be presented collectively in Section 4.3.5 based on the findings from the three substantive inquiries. They will also be used for a construct appraisal in light of the test purpose, followed by a decision-making for the validity investigation.

4.3.2 Test-taking processes and strategies

The first inquiry is to analyze all of the verbal reports collected in ConCloze 2–4. The processes and strategies already constructed for verbalizations to one item will be verified as to whether they are still applicable to a larger number of items. Investigating the applicability of the processes constructed can be useful because their accuracy could be argued as high if they need little adjustment for cross-item application. Describing the underlying processes accurately would then confer their utility, which in turn indicates (a) accuracy of the test-score interpretations, and (b) likelihood of accounting for other task engagements in the universe of admissible observations. Representing multiple sets of task content in testing the construct proficiency, their power of generalization could thus be the plausibility of the validity argument.

The procedure of the current inquiry is similar to that illustrated in Figure 34 (pages 125ff.). A main distinction is to add more details to the verbalizations. For example, single quotations (‘’) are used for demarcating the words and phrases that are likely to be read directly from the item. Square brackets ([]) are also added for notes. An annotated instance would be *It’s it’s it’s not familiar to me. [Line 5, only one word, right after the KWIC position, with KWIC inserted] [Option A] ‘avian avian’ drugs’? I’m not, not, I cannot the meaning! haha [laughing]* (Aaron, ConCloze 2 Item 4). Adding more of such details is intended to make the segments of verbalization more meaningful when migrated individually to the database. Owing to this modification in transcription and annotation, verbal reports to Item 2, which have been analyzed in ConCloze 2–4 (pages 128–145), are also re-processed for consistency in the current analysis. The transcriptions of the reports are provided in Appendix 6 (page 390) and Appendix 7 (page 408).

The processes and strategies are enumerated in Figure 72 below. Based on 60 verbal reports (12 respondents \times 5 items), the total number of segments identified is 2,487. A most distinctive process constructed is *Testing compatibility of a given word in context*. This process outnumbers all the other processes and strategies consistently across the items, taking 34% of the segments (836 of 2,487). The process is also identical to the preliminary one previously constructed in ConCloze 2–4 (cf. page 136). An interpretation is that ConCloze examinees engage in the test tasks by actively checking if an option fits the concordance-based context provided. This means that each option word could contain semantic components, which are to be matched with the word information related to it

that is embedded uniquely in the concordance prompt. Accordingly, an inference can be that proficiency in vocabulary use is tested in a bipartite manner during ConCloze engagement.

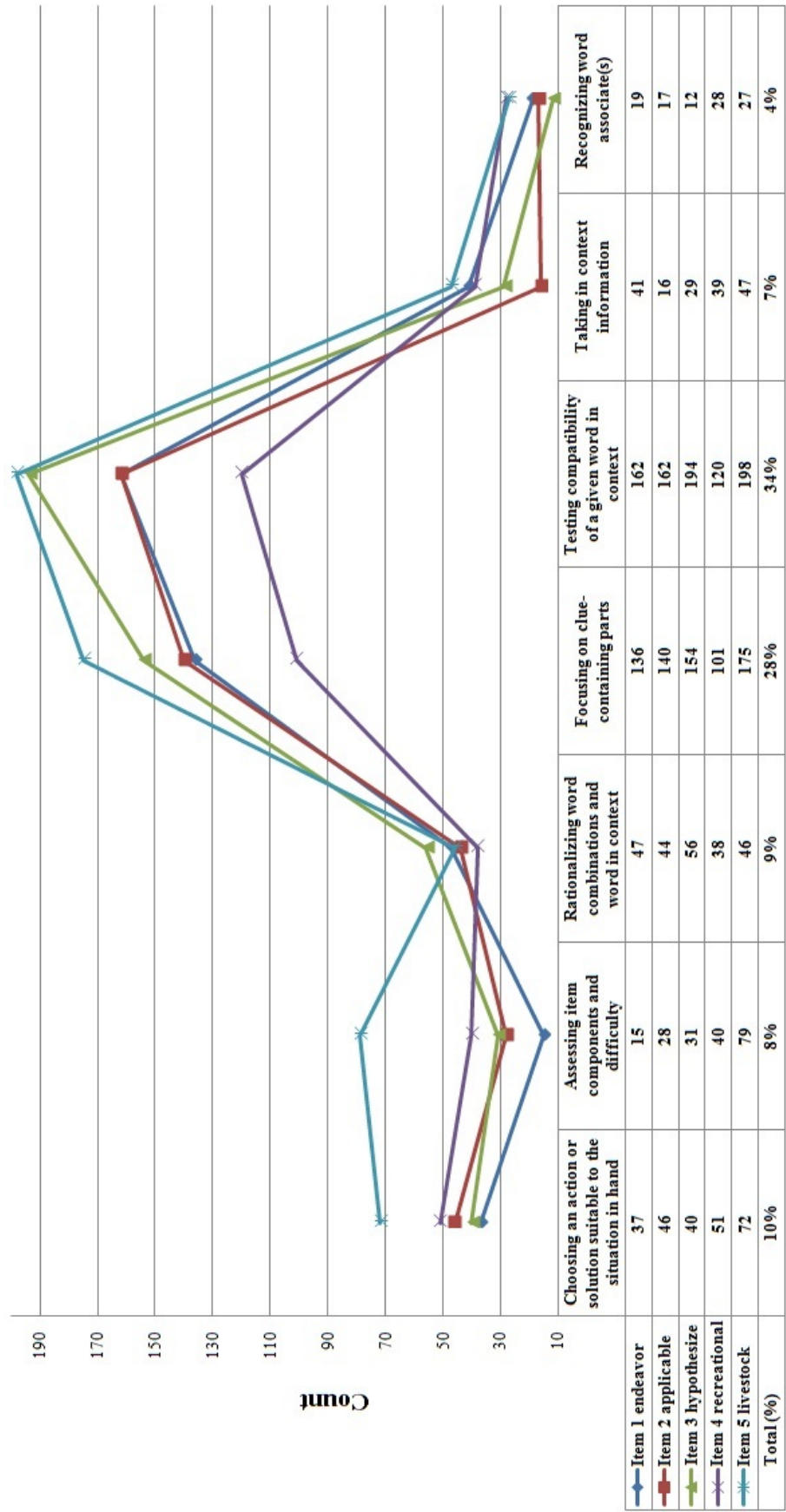


Figure 72 Enumerating test-taking processes and strategies

In addition to the prime process, four other processes from ConCloze 2–4 are also found applicable in all the items of Figure 72: the strategy *Focusing on clue-containing parts* (28%), the process *Rationalizing word combinations* (9%), the strategy *Assessing item components and difficulty* (8%), and the process *Recognizing word associate(s)* (4%) (cf. pages 134ff. for those originally constructed). An inference is that the examinees test compatibility of a given word in context, usually by allocating the capacity of their working memory to specific concordance-derived information that they believe contain significant clues. This has been similarly argued for as a strategic allocation of on-line concentration in ConCloze 2–4 (page 136). Current examples include Figure 73 below, in which Aaron reflected on focusing on one line (i.e., Line 5) which he had found particularly helpful. Given the prevalence of this strategy and of the prime process, it may be argued that they are core processes mobilized when the examinees engage with ConCloze tasks. Because they do not need rewording, the processes seem accurate in describing the substantive aspect of task processing.

1 It is not unreasonable to _____ a heightened sense of territoriality
 2 knowledge will lead us to _____ different authorial and narrative audiences
 3 $p < 0.001$). # Alternatively, one could also _____ that adolescents with high consumption
 4 reading of the literature, they _____ ed and found three distinct clusters
 5 1992). In other words, we _____ ed teenage sexual behavior would be
 6 happier than single people and _____ ed that marriage may draw persons
 7 behavior. Notably, parental smoking is _____ ed to demonstrate pro-smoking norms and

All the lines above miss the same word. Which of the following should be that word?

- ☐ A formulat(e)
☐ B hypothesiz(e)
☐ C suggest
☐ D verify(/i)

Name	Com Unit	Key	Process Level	Process/Strategy	Hint	Verbal Report
Aaron	27	hypothesize	Concurrent	Focusing on clue-containing parts	Non-linear segmenting	Researcher: Alright, ok, er. Which word do you pay particular attention to or do you read the whole er back part? Aaron: Er, I think I read one sentence after another but when I read the fifth question fifth line [i.e., Line 5] I think [P] Researcher: The fifth line? Aaron: Yeah, the fifth line [i.e., Line 5] [P] helps me [P] maybe [Option C] 'suggest' is appropriate and I haven't read the last two sentences. [P] Researcher: Alright. Thank you so much. The next one please.
Aaron	28	hypothesize	Retrospective	Choosing an action or solution suitable to the situation in hand	Explanation, decision-making	Researcher: Alright, ok, er. Which word do you pay particular attention to or do you read the whole er back part? Aaron: Er, I think I read one sentence after another but when I read the fifth question fifth line [i.e., Line 5] I think [P] Researcher: The fifth line? Aaron: Yeah, the fifth line [i.e., Line 5] [P] helps me [P] maybe [Option C] 'suggest' is appropriate and I haven't read the last two sentences. [P] Researcher: Alright. Thank you so much. The next one please.

Figure 73 Strategy 'Focusing on clue-containing parts' and process 'Choosing an action or solution suitable to the situation in hand'

Apart from the core strategy, Figure 73 above also displays another strategy: *Choosing an action or solution suitable to the situation in hand*. Newly constructed in this inquiry, the strategy is marked for 10% on average of all the verbalization segments (246 of 2,487). The criterion is when a verbalization contains explicit decision-making in relation to the task being dealt with or when an action or a series of actions is justified retrospectively. For example, in Figure 73, Aaron explained that he had read up to the

fifth concordance line before deciding that reading Lines 6 and 7 was unnecessary. This strategy seems to accord with the notion of test management such as self-monitoring (cf. Cohen 2012 for a classification of test-taking strategies). Accordingly, it can be argued that meta-cognitive planning and decision-making is inherent in ConCloze engagement alongside language-related processes.

In addition to the strategic decision-making, the other new process constructed in Figure 72 (page 230) is *Taking in context information*. The process is marked for 7% on average of all the verbalization segments (172 of 2,487). The criterion is when the respondent appears to accumulate task content before attempting to select an option. This is usually determined by analyzing the flow of the verbalization such that selecting the option(s) emerges later. Examples include Figure 74 below, in which Lulu began by reading the first two concordance lines and all of the options. After moments of possibly deep processing as evidenced by pauses, she then expressed her doubt as to whether any of the given options could be the answer. This means that before reading the options, Lulu must still be taking in information of Lines 1 and 2, rather than seeking to test their compatibility yet.

1 It is not unreasonable to _____ a heightened sense of territoriality
 2 knowledge will lead us to _____ different authorial and narrative audiences
 3 $p < 0.001$). # Alternatively, one could also _____ that adolescents with high consumption
 4 reading of the literature, they _____ ed and found three distinct clusters
 5 1992). In other words, we _____ ed teenage sexual behavior would be
 6 happier than single people and _____ ed that marriage may draw persons
 7 behavior. Notably, parental smoking is _____ ed to demonstrate pro-smoking norms and

All the lines above miss the same word. Which of the following should be that word?

- ☐ A formulat(e)
☐ B hypothesiz(e)
☐ C suggest
☐ D verify/(i)

Name	Com Unit	Process Key	Process Level	Process/Strategy	Hint	Verbal Report
Lulu	1	hypothesize	Concurrent	Taking in context information	P, E	Next one, [P] [Line 1] 'It is [E] not [E] unreasonable [E] to [E] [KWIC position here] [P] a [E] heightened [E] sense of er [E] territoriality [E]', [P]
Lulu	2	hypothesize	Concurrent	Taking in context information	P, E	[P] [Line 2] 'knowledge will [IA] lead us to [P] [KWIC position here] different authorial and narrative audiences' [P]
Lulu	3	hypothesize	Concurrent	Taking in context information	P, E	[P] [Option A] 'formulate', er [Option B] 'hypothesize', [Option C] 'suggest', [Option D] 'verify' [P] er [P]
Lulu	4	hypothesize	Concurrent	Testing compatibility of a given word in context	P, E, Inserting options	[P] [Option A] 'formulate', er [Option B] 'hypothesize', [Option C] 'suggest', [Option D] 'verify' [P] er [P] Well, I feel like none of this [options] fits the sentences, haha [laughing]

Figure 74 Process 'Taking in context information'

The following is an implication of the findings presented thus far. In ConCloze 6 (page 224), a composite meaning is proposed for ConCloze scores. The competence domains involved include knowledge of word association and lexical-semantics, superordinated by knowledge of synonymy and collocation. They are mobilized through the process of testing the compatibility of the lexical semantics with the clues of

concordance-based information. In this inquiry, the core processes and strategies are found to operate as substantively expected and consistently across items. A fresh insight is that the examinees are also likely to manage the test tasks metacognitively by strategic decision-making and allocation of their short-term processing capacity. Most importantly, evidence in this inquiry is a first piece of substantive evidence to address the inadequacy mentioned in the rationale (page 226). Thus far in this study, the processes and strategies constructed for task engagement seem to be applicable to two separate scales of performance assessment: a single-task preliminary (Section 3.3.4) and a multiple-task follow-up. Given the inter-item consistency in test-taking processing, the score interpretations could be claimed to have been granted a power of generalization and are likely valid for the ConCloze item type.

4.3.3 Substantive contrast of sample

A second inquiry for substantive fine-tuning is to maximally contrast a sample of verbal reports. In contrasting them, linguistic analyses of item components will also be incorporated so as to find out if the language features at face value are reflected in actual engagement. Including face-validity evidence can be useful because the insights based on the verbalizations may or may not fully reflect the respondents' proficiency levels tested. For example, the test tasks in ConCloze 2–4 (page 130) and Section 4.3.2 (page 233) are found during concurrent verbalizations to often invoke deep processing, which suggests some elusiveness in interpreting score meaning (cf. Pressley & Afflerbach 1995: 2ff. for challenges in analyzing verbalizations such as occasional elusiveness). By contrast, analyzing task content with the answers known is a top-down perspective; sources of information such as corpora can be sought and complement such bottom-up insights as those from the verbalizations. This means that combining both of the perspectives can enrich the score interpretations in the way that may not be achieved otherwise.

The inquiry process can be summarized in Figure 75 below, the steps of which will be consulted along this inquiry. First of all, Item 4 (target word: *recreational*) is chosen purposely from ConCloze 2–4 because the respondents should have been much at ease with verbalizing the test task after engaging with Items 1–3. Being at ease with the verbalization tasks means that the sampled weak respondent (described later) would not be disadvantaged in expressing himself. Item 4 together with a collocation report is

illustrated in Figure 76 that follows, in which collocates of the target word *recreational* are identified at nine words on either side—the maximum span of collocation computable at COCA. The aim of deriving the collocation report is to replicate as closely as possible the processing of significant words of association in the prompt by the examinees. Twelve words in the prompt are found to also rank collocationally high in the corpus, e.g., *activities* (ranked 1st) and *use* (ranked 2nd). On the one hand, the concomitant occurrence can be anticipated because the concordance lines have been retrieved from COCA, just like Gardner & Davies’s (2014) AVL, which is similarly based on COCA (cf. pages 71ff. for the spec of the word list used in this study). Yet, in terms of corpus empiricism, these collocates in the prompt could be argued as naturally expectable near *recreational*. An inference is that the collocates may also activate their links with *recreational* in the mental lexicon and thus serve as context clues during ConCloze engagement.

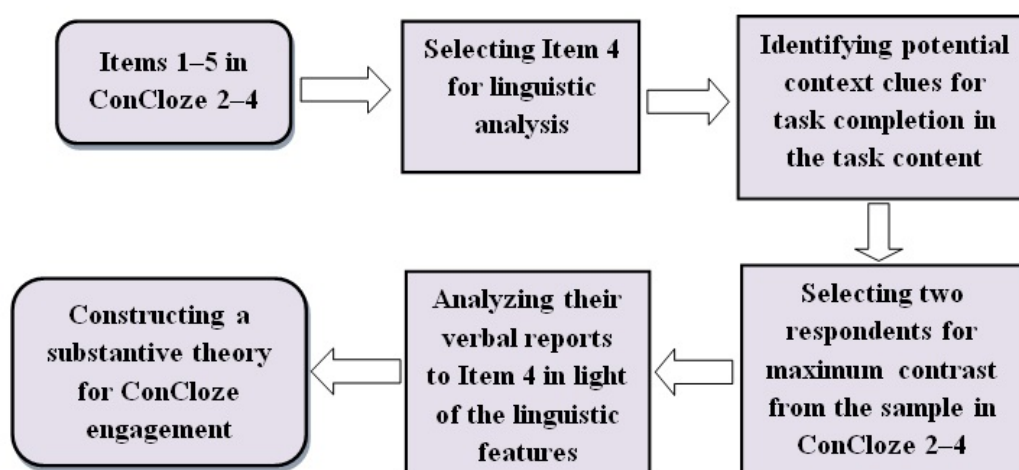


Figure 75 Contrastive analysis with face-validity evidence

1 commitments in their lives such as family, employment, and other activities and hobbies that capture their interest. The college players
2 frequent, use. If their use of these drugs is often and associated with leisure activities, then it is likely that
3 economic base of the Andean countryside shifted from agriculture to attractions. Tourism intensified after devaluation of the bolivar in 1983
4 art activities, and the green token represented fifteen minutes of computer time on a typical classroom computer (i.e., Apple IIe
5 defined polydrug use as the use of 3 or more drugs in the past 90 days (meth-amphetamine, cocaine, crack, MDMA
6 on how well the library was meeting users' educational and needs so improvements could be made. They agreed to complete
7 to the United States, where most bicycles sold are for use, most of the 105 million bicycles sold each year worldwide

All the lines above miss the same word. Which of the following should be that word?

A active	B distracting		C enjoyable		D recreational	
Rank	Word	Rank	Word	Rank	Word	
1	activities*	8	opportunities	19	users*	
2	use*	9	sports	25	competitive*	
3	fishing	10	park	26	educational*	
4	commercial	11	areas	31	boaters*	
5	facilities	14	athletes*	43	leisure*	
6	drugs*	17	vehicle	105	cocaine*	
7	drug*	18	purposes	121	marinas*	

* Found in the concordance prompt of ConCloze 2–4 Item 4

Figure 76 ConCloze 2–4 Item 4 with a collocation list

The next step in Figure 75 (page 236) is to identify potential context clues that could be systematically used during task engagement. Clue identification is performed by incorporating the present collocates in Figure 76 into a clue map in Figure 77 below. The clues are grouped into three partially overlapped categories: (a) collocates, (b) clues describing the target word *recreational* via their lexical content (i.e., word meaning, usually compositionally), and (c) those describing the target word via their lexicogrammatical structure (i.e., position of the target word with the lexical content in the line also considered). The clues are intended for expansive reading: from individual words of collocation on the left to a convergence of lexical-semantics and syntax on the right. For example, in Line 1 of the item prompt (cf. Figure 76 above for the concordance line), the expression *capture their interest* is part of the predicate in the relative clause modifying the noun phrase *other recreational activities and hobbies*, which also contains the target word. The expression refers to something interesting and exciting and can convey a positive meaning of the target word. Accordingly, it is the compositional meaning that may serve as a hint at the target word for the examinees. Another example is in Line 6, where the target word is used in the parallel structure *educational and ____ needs*. An examinee may infer that this target word, when considered amidst the co-text, must involve another purpose in use of a library other than an educational use. Thus, the lexical meaning interpreted in the syntactic structure could be argued as a clue hinting at the target word *recreational*.

Recreational		
Collocates: <i>activities, athletes, boaters, cocaine, competitive, drug(s), educational, leisure, marinas, use, user(s)</i>	Descriptions via lexical- semantic content: Line 1: <i>capture their interest</i> → Entertaining Line 2: <i>leisure activities</i> → Not job-related; pastime	Descriptions via lexicogrammatical slot: Line 3: <i>comparisons between competitive and ____ athletes</i> → The other category of athletes? Line 4: <i>marinas [...]</i> ____ <i>boaters</i> → Word to describe boaters at marinas Line 5: ____ <i>drugs [...]</i> <i>amphetamine, cocaine</i> → Word to modify illegal drugs Line 6: <i>library was meeting users' educational and ____</i> <i>needs</i> → Purpose in library use if not for education Line 7: <i>in the U.S. [...]</i> <i>most bicycles [...]</i> ____ <i>use</i> → Main use of bicycles in America: To commute? To work out? To relax?
Thematic summary: A word used for describing a fun activity done as pastime		

Figure 77 Potential clues in Item 4

It is worth considering that world knowledge is often called for in the categories *Descriptions via lexical content* and *Descriptions via lexicogrammatical slot* of Figure 77. This is usually in the same lexical-semantic baggage carried by individual words and their combination; for example, those who never use a library may be oblivious of multipurpose use of a library as interpretable in Line 6. Another example is in Line 5, where the grammatical slot ____ *drugs* seems to demonstrate that the target word can modify the noun *drugs*. Irrespective of this structural information, it is rather the semantic baggage of the words *amphetamine* and *cocaine* in the co-text, for instance, which may help to determine that the target word, yet unknown to the examinees at the beginning of the task, can be used for describing narcotics. For this reason, such lexical-semantic baggage could be deemed semantic prosody which words carry and which the examinees may be cognizant of through their world knowledge and language knowledge acquired (cf., e.g., pages 7, 51, and 59 for semantic prosodies as a language component contained in a concordance; cf. also Hargreaves 2000: 213ff.; Hoey 2000: 231ff. for semantic prosody as a kind of collocation that language learners need to internalize).

In light of the world knowledge required for processing lexical-semantics, it is worth stating that categorizing the clues here is for analytical purposes. In real-time task engagement as in ConCloze 2–4, no respondents are found to systematically specify the domains of competence they could claim they were using during task engagement. For example, none of the sampled respondents would metacognitively verbalize “OK, now I am going to use my background knowledge about this word.” Given this, it may be argued that ConCloze examinees could be unaware whether they are utilizing their world knowledge in unlocking the lexical-semantic baggage or their lexicogrammatical knowledge in drawing inferences from the lexical–syntactic clues.

In ConCloze 1–5 (e.g., page 98), lexical-semantic knowledge, particularly core components in meaning, is argued to be mobilized in task processing. In light of the current mapping of context clues, a top-down processing can be modeled in Figure 78 below. The clues are categorized conceptually into four types: (a) word meaning (including semantic prosody and connotation), (b) compositional meaning and subject matter (including world knowledge), (c) collocation, and (d) structure. It is worth highlighting that collocation is assigned to the class *formal representation* because it is habitual co-occurrence closely tied with word forms rather than lexical meaning only. For example, *cool* and *cold* can be considered near-synonyms, but would not be

interchangeable in *cool rack* and *cold call* (cf., e.g., Hoey 2000: 232ff. for similar restrictions on using ‘chilly’). Then the clues could likely assist the examinees in conceptualizing a lexical representation of the target word, which would eventually match with the meaning or core components in meaning of one of the option words.

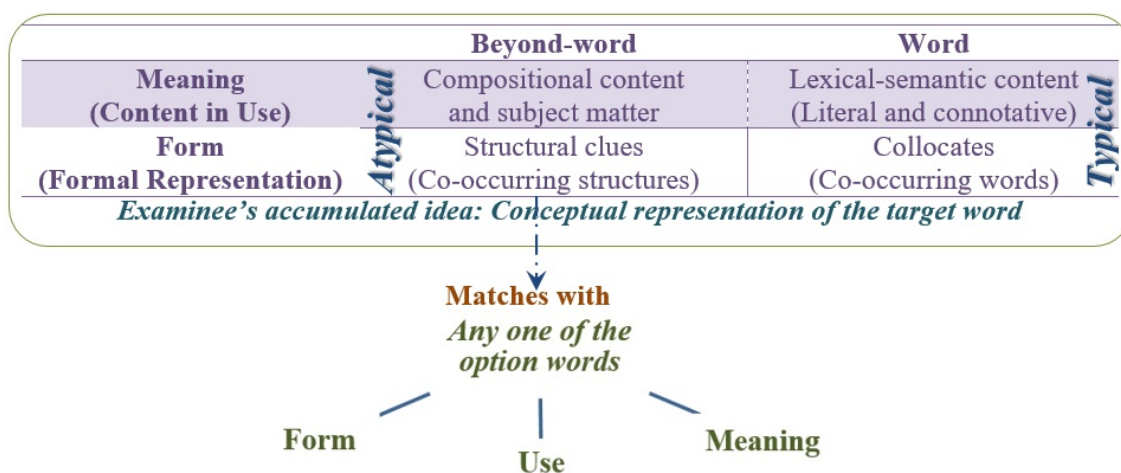


Figure 78 Conceptualized ConCloze processing

In addition to proposing the engagement mechanism as triggered by various context clues, the clues are also systematized into typically and atypically co-occurring ones in Figure 78. The former co-occurrence is habitual, relying on the normal likelihood of finding the clues in the vicinity of the target word, e.g., the present collocates of *recreational*. The atypical co-occurrence is text-specific and is processed and learned on site from the prompt, determined by particular discourse structure and author styles, as well as the subject matter requiring world knowledge when interpreting. For example, the structural parallelism *comparisons between competitive and recreational athletes* is identified as a clue in Line 3 of Item 4 (page 239). Depending on the attributes of the original text, the parallelism seems unable to be expected as habitually co-occurring with *recreational*.

Given that the potential clues contained in the item prompt have been identified, the next step in Figure 75 (page 236) is to seek a pair of respondents from ConCloze 2–4 for maximum contrast. Halle and Igor are selected from ConCloze 3 because they could be deemed the sharpest contrast in general English proficiency in Table 21 (page 112). When signing up for the test, Halle self-reported an IELTS result of 8.5 obtained in 2011

whereas Igor scored 5.5 in 2010. The transcriptions to their verbal reports are provided in Appendix 7 (page 408) and will be analyzed in turn. Overall, Halle's task engagement is marked by more pauses closer to the beginning of the report than near the end. In ConCloze 2–4, such pauses are interpreted as episodes of deep processing, potentially indicating intermittent evaluations (page 130). A current example is in Line 1, where she did not focus on any particular phrase but read it once with pauses seen. After that, she read all the option words with more pauses seen, expressing uncertainty as to the exact meaning of Option A *avian*. Then she read the line again, now with Option C *recreational* inserted. Afterwards, she continued reading the rest of the lines in the item prompt, mostly once, with occasional pauses and with the chosen option word inserted along.

Apart from the momentarily deep processing, another observation is that Halle might be using interspersed 'phrasal checks'—reading one particular phrase of a certain length but with the chosen word inserted. Namely, she intensely worked on the first line she came across, decided on one option word, and systematically applied that decision to the rest of the concordance lines. This manner of progressing with the task could be deemed ancillary to the main process *Testing compatibility of a given word in context* and accord with the primary strategy *Focusing on clue-containing parts* (e.g., pages 138 and 230). The phrases Halle checked with can be illustrated in Figure 79.

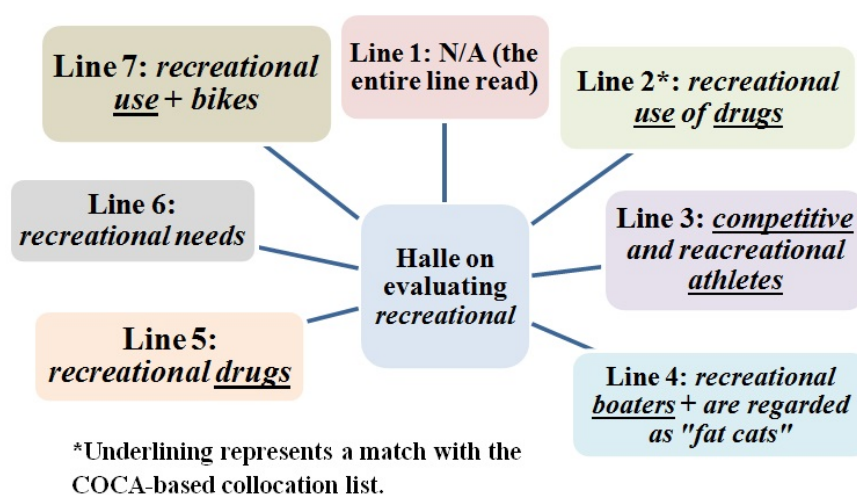


Figure 79 Halle's phrasal checks

Considering Halle's use of the primary strategy *Focusing on clue-containing parts*, it is worth pointing out that most of the phrases in Figure 79 encompass some of

the COCA-based collocates in Figure 76 (page 237, underlined here for emphasis in Figure 79). In fact, *vehicle* (rank 17th in Figure 76) can also be considered a hypernym of *bikes* as verbalized by Halle and *purpose* (rank 18th in Figure 76) a near synonym of *needs*. This seems to suggest that the phrases Halle focused on contain naturally occurring collocates. An inference is that she might recognize these phrases as significant for the answer she had chosen and allocate her processing capacity accordingly. On this account, a conclusion could be that high performers in ConCloze may systematically check the phraseological quality of their answers during task engagement.

With regard to the processing of a low performer, Igor started similarly to Halle by reading Line 1 once. In the second reading, he also verbalized nearly all of the options in place of the KWIC blank. He then expressed his uncertainty in his knowledge of vocabulary—potentially relating to the meanings of the option words—and evaluated that the word fitting the blank should be an adjective. Igor then continued by deciding between Option C *mechanical* and Option D *recreational*, which culminated in the latter choice. After he continued to Line 2, with pauses seen and with partial repetitions, the chosen word was inserted. Upon identifying that he had settled on an option, the researcher started to interview him retrospectively. Only then did he appear to overturn his decision, saying that Option D (which is in fact the key) was wrong. He read Line 2 again, and said that the blank would require a past participle.

The description above is only part of Igor's engagement with Item 4, but a general observation is that Igor seemed to focus on fewer concordance lines than Halle did and to be unaware that all the option words provided are adjectives. In fact, the option words all bear adjectival morphological markers: *-ian*, *-al*, and *-y*. Moreover, his verbal report is marked by signs of uncertainty throughout, such as pauses, mumblings and the sound *er*. Potentially driven by an inability to complete the task meaningfully, he might compensate for his lack of knowledge of the meanings of some of the option words by processing immediate local constituents instead. In other words, upon partial knowledge, localized processing could be performed for task completion: Igor reflected in Figure 80 that he processed a surrounding structure in order to figure out the word class of the KWIC.

Igor: [Chuckle] [P] Yeah, this one, OK, I choose, [P] [Line 5, in phrase, from right before the KWIC position, with KWIC inserted] 'more [P] [Option D] 'recreational' drug,' [P] [Option C] 'mechanical', [IA] I think, OK, should be D [Option D 'recreational'], I guess. [chuckle]

Researcher: Why?

Igor: This one, I just er see the structure of the sentence, and I see some [P] some vocab missing

Researcher: For example?

Igor: Yeah.

Researcher: What do you mean, 'structure'?

Igor: Yeah.

Researcher: Of the sentence?

Igor: Yeah, this one if I don't know about the the meaning of the vocab, so I will s see the structure. Like [P] sometime[s] verb, sometime[s] adjective missing. So, this one you need adjective.

Researcher: Uh-huh.

Figure 80 Igor's retrospection

In addition to observing Igor's local concentration, his phrasal checks seem to reflect a similar picture. Summarized in Figure 81 below, three of his seven phrasal checks are found to involve structural analyses, rather than the meanings of the option words or semantic relations among them. Furthermore, Igor repeated verbalizing parts of a few concordance lines, but his repetitions seemed unsystematic. On the one hand, he did not read the whole concordance prompt to the extent that his phrasal checks cannot be established from the unread lines (viz. Lines 3, and 6–7). Yet, his omission of verbalizing a few concordance lines despite a chance of doing so seems to indicate his lack of motivation in completing the task meaningfully, if not an inadequacy of construct-relevant ability. It thus may be inferred that Igor strategically focused on local, most likely grammatical, clues rather than the clues of lexical-semantic content and compositional meaning on a wider scope of the concordance prompt.

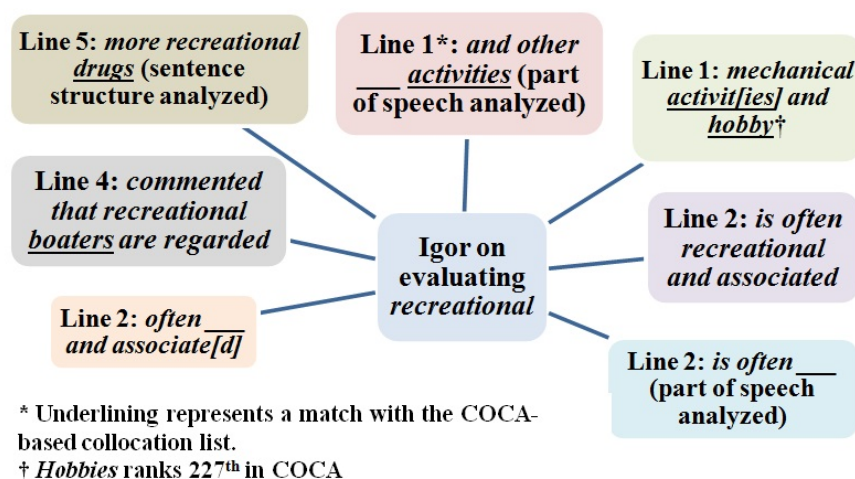


Figure 81 Igor's phrasal checks

Earlier in the thesis (pages 39, 39, and 44), ConCloze is theorized to serve as a proficiency test in the sense that examinees with a higher proficiency level would stand a chance of scoring higher in the test than those of lower construct proficiency. In light of the different styles of Halle's and Igor's task engagement, a substantive processing can be concluded as follows. First, Halle systematically dealt with one concordance line and checked her choice in the rest of the lines. By contrast, Igor seemed to search in an unorganized way for fortuitous clues. While both of them arrived at the correct answer, their engagements differ clearly in efficiency. In fact, it is also worth highlighting that just because an examinee has a low level of the construct proficiency does not mean that the examinee has no proficiency in the competence tested altogether (page 45). Even though potentially at different levels as inferred from their IELTS scores (page 241), both of the respondents may be deemed to have the construct proficiency—which has thus far been defined in light of the test purpose as that in academic vocabulary use (pages 224f.).

Secondly, Halle's phrasal checks appear to systematically encompass some of the natural collocates in the prompt whilst she was reading the concordance lines. By contrast, Igor seemed to read broadly and rely on pondering over grammatical clues, particularly in the form of syntactic frames contiguous to the KWIC blank. Illustrated in Figure 82 below, an inference is that Halle accumulated global context clues with local foci on collocates whereas Igor sought to look primarily for local syntactic clues. Thirdly, both of them read concordance lines in full and in part, suggesting that the lexical-semantic content of individual words and of them in combination is likely to be processed during

task engagement. Lastly, neither of them seemed to use world knowledge explicitly. Given this, world knowledge in its pure form and as modeled in Figure 78 (page 241) can only hold marginal construct relevance. Also, mobilized by the low performer only, knowledge of grammatical structure may likewise be assigned a peripheral role in ConCloze-score meaning. In sum, it may be argued that the high performer's word knowledge could be more profound than that of the low performer and thus allow her to check a meaningful compatibility between the given option words and the concordance-derived clues.

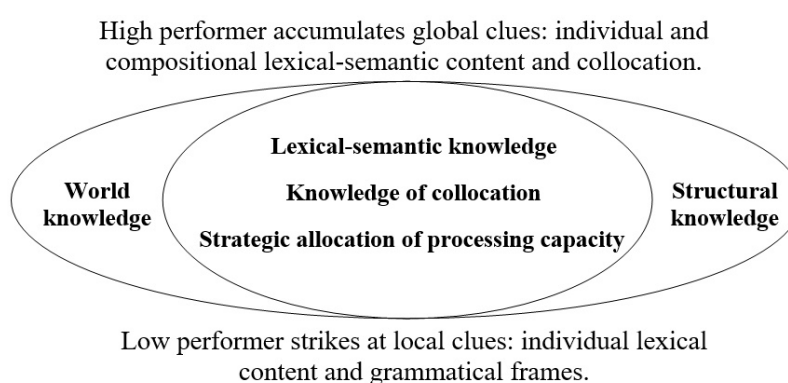


Figure 82 ConCloze competence based on a contrastive sample

4.3.4 Modified constructed-response format

The last inquiry for substantive fine-tuning is to devise a modified constructed-response format. The aim is to elicit another set of verbal reports and investigate if the underlying processes are comparable to those applied in the previous section 4.3.2. Exploring the comparability of the processes in the new verbalizations can be useful for two reasons. First, the current item format is different from the selected-response format administered in ConCloze 1–6. Generally, processes underlying task engagement could vary depending on test methods (e.g., Buck 1991 for effects of differing research methods on performing in listening tasks). Finding core processes amidst the different types of expected response can avert test-method bias and would confer generalizability for the substantive-validity theory developed (cf. D'Agostino 2005 for threat of mono-method bias towards construct validity). Secondly, the test used in this inquiry is paper-based, a delivery mode different from the internet-based delivery in ConCloze 1–6. Change in the mode of test presentation may change task processing, which would call for a construct

redefinition (Bachman 2000: 9). Accordingly, comparability of the processes constructed can also address delivery-mode bias and would likewise confer wider applicability of the substantive theory for this validity study. On these accounts, finding substantive comparability between different test methods and test modes would strengthen the test interpretations and make a case for test usefulness.

The inquiry process can be summarized in Figure 83 below, most steps of which are similar to those in ConCloze 2–4 (pages 125ff. and 133f.). A distinction lies in the current use of an intervention technique, in which the respondents work on the first part of each item whilst being considered as to whether the second part should be given (cf. Anderson *et al.* 1991: 47 for a similar technique during verbalization elicitation). Illustrated in Figure 84 that follows, the first part consists of the item prompt and the question stem, and the second part has the options. The reason for preparing the option sheet is that completing constructed-response items is generally more onerous than doing selected-response items, *ceteris paribus*. Uncertainty still exists regarding if a constructed-response ConCloze in its strong form is ever feasible. Hence, having an option sheet ready—albeit not informing the respondents in advance—means a contingency plan for an effective elicitation of verbal reports. Appendix 5 (page 388) offers items of the current inquiry, and Appendix 8 (page 479) the verbal reports.

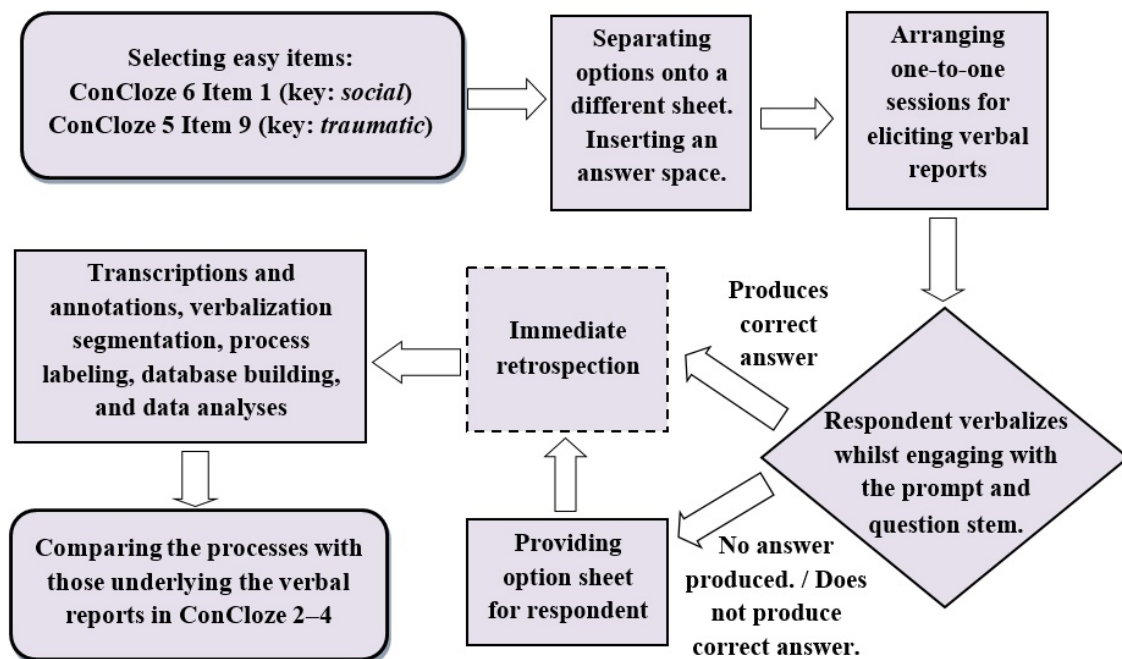


Figure 83 Modified constructed-response format: From administration to analysis

1 and therefore it is his _____ childhood experience that is to
 2 comparative framework and not other _____ contexts such as Rwanda or
 3 her rape but the enduring _____ effects of it on her
 4 has been fully experienced, the _____ event moves from the present
 5 around two points: first, that _____ events have usually been those
 6 was clear that the woman's _____ experience during World War II
 7 of coping with and expressing _____ issues and experiences. The individual

All the lines above miss the same word. What should that word be?

- A anxious
- B conditioned
- C disagreeable
- D traumatic

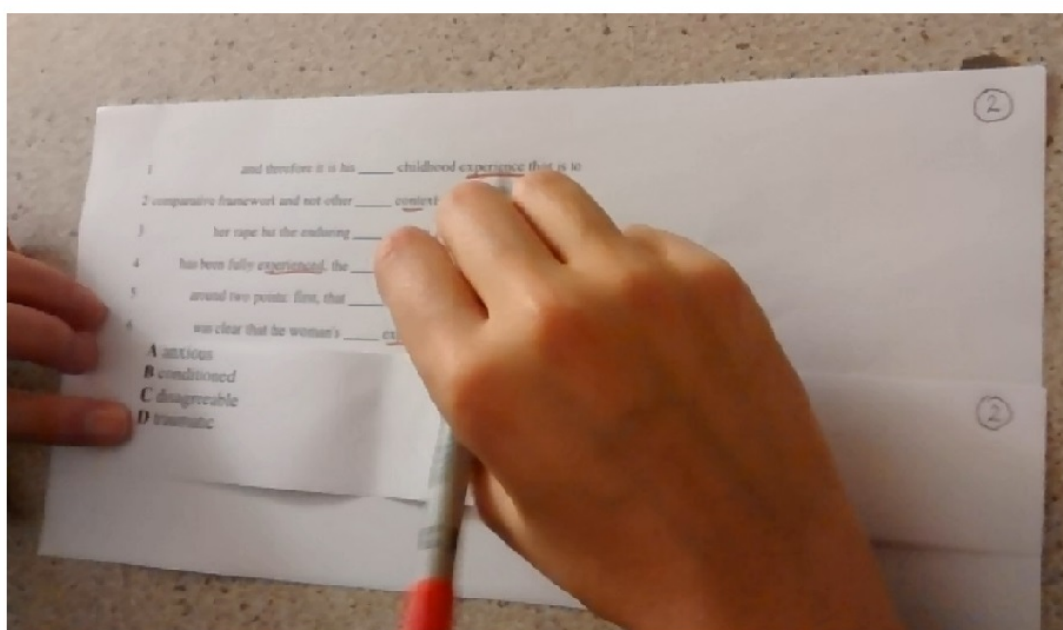


Figure 84 A two-part ConCloze item

The details of the respondents in this inquiry are shown in Table 43 below. For heterogeneity, their education levels and L1s are deliberately diversified as these two variables are found to be potential surrogates for the intensity and the extent that the proficiency tested in ConCloze is acquired (pages 198 and 203f.). A first step in conducting the inquiry is to investigate two issues in their task engagement: whether the

paper-based mode is feasible and whether the constructed-response format is feasible, which will be discussed in turn.

Regarding the feasibility of the test mode, the criterion is general convenience in dealing with the paper-based test. It is found that no sign could be observed indicating that either of the respondents struggled to engage with any of the items on paper. Nor did they mention any difficulty tackling the paper-based tasks. In fact, the respondents appeared entirely familiar with handling the test with paper-and-pen. For example, Figure 84 above demonstrates that Nina moved the option sheet closer to the prompt and on top of the question stem without being asked to. On this account, it seems that the paper-based mode is unlikely to cause response invalidity in this inquiry and would be feasible for general testing purposes.

Table 43 Details of ConCloze 7 respondents

Name*	Age	Level	L1	Latest Standardized Test Reported	
				Year	Result
Maya	23	Undergraduate	Punjabi	N/A	N/A
Nina	22	Taught postgraduate	Portuguese	2014	IELTS 7.0

* All pseudonymized

With regard to feasibility of the constructed-response format, the administration in Figure 83 (page 247) is designed such that the option sheet can be handed out when the respondent is unlikely to produce a plausible answer or no longer attempts to deal with the test task. Exemplified in Figure 85 below (bold typed), this is when more pauses could be observed and the respondent signaled a stop in task engagement based on the item prompt. The criterion for assessing feasibility of the format is when the respondents engaging with the constructed-response ConCloze can produce a correct or nearly correct answer. Figure 85 demonstrates that none of their attempts at producing answers invokes the target words. The most positive effort seems to belong to Nina in Item 1, who arrived at *ethnic* in lieu of the key *social*.

Given the respondents' failure to produce correct answers, it is also worth considering that the items administered are deliberately chosen on the basis of their relative easiness—Item 1 with the original item estimate of -1.49 , and Item 2 -0.98 . Because an item estimate of between -1 and 1 can be deemed moderately difficult, the

original item estimates can be interpreted such that the items are very easy and easy, respectively. This implies that the respondents' failure to arrive at the correct answers could not be attributed solely to their low level of the ConCloze proficiency. Rather, removing the options from these items could drive the difficulty, causing a mismatch between the proficiency level and task difficulty. Accordingly, no sufficient evidence is obtained in arguing for the feasibility of the constructed-response format in its strong form. It seems that a modified one would be more appropriate for general test use.

Maya: [Line 2 again immediately, in word, right before the KWIC position] 'behaviors
['and' not verbalized] [KWIC position here]' [P]

Researcher: Keep saying. Whatever you are thinking.

Maya: Erm, I don't know that question [i.e., the concordance line]. [P] So, it can't it
actually can't be the the word the same words. [P] *Erm.* [P] *Actually dunno.* [P]

Researcher: OK. Now, if you have four choices here [the answer options now placed
before Maya],

Maya: Erm. [P] [Looking at the answer sheet.]

Researcher: Keep saying. Whatever you are thinking or reading.

Item No.	Source	Item Estimate	Key (COCA Rank)	Item Features	Answer Attempted	
					Maya	Nina
1	ConCloze 6 Item 1	-1.49	<i>Social</i> (4th)	Seven words on either side, Four semantically unrelated options	<i>the</i>	<i>ethnic</i>
2	ConCloze 5 Item 9	-0.98	<i>Traumatic</i> (1241st)	Five words on either side, Four semantically related options	N/A (with <i>events</i> and <i>experience</i> written near the prompt, potentially as theme)	N/A

Figure 85 Intervention with an option sheet, and answers attempted

Apart from the feasibility evaluation, the next step in this inquiry is to investigate the underlying processes. Considering the infeasibility of the strong form of the constructed-response format, this inquiry treats each verbal report as a continuous flow of task engagement irrespective of when the option sheet is given to the respondents. Summarized in Figure 83 (page 247), the verbalizations are processed in a similar way to the steps taken in ConCloze 2–4. Based on four verbal reports (2 respondents \times 2 items), the total number of verbalization segments identified is 108. The substantive processes are presented in Figure 86 in segment count and in Figure 87 that follows in percentage.

<i>Process/Strategy</i>	<i>ConCloze 2–4 (Selected-response)</i>	<i>ConCloze 7 (Modified Constructed-response)</i>	<i>Process/Strategy</i>
Choosing an action or solution suitable to the situation in hand	246	12	Choosing an action or solution suitable to the situation in hand
Assessing item components and difficulty	193	0 (N/A)	
Rationalizing word combinations and word in context	231	13	Rationalizing word combinations
Focusing on clue-containing parts	706	14	Focusing on clue-containing parts
	0 (N/A)	2	Testing compatibility of a retrieved word in context
Testing compatibility of a given word in context	836	12	Testing compatibility of a given word in context
Taking in context information	172	35	Taking in context information
Recognizing word associate(s)	103	9	Recognizing word associate(s)
	0 (N/A)	11	Retrieving possible words
<i>Total</i>	<i>2487</i>	<i>108</i>	<i>Total</i>

Figure 86 ConCloze 2–4 and 7 processes and strategies in number

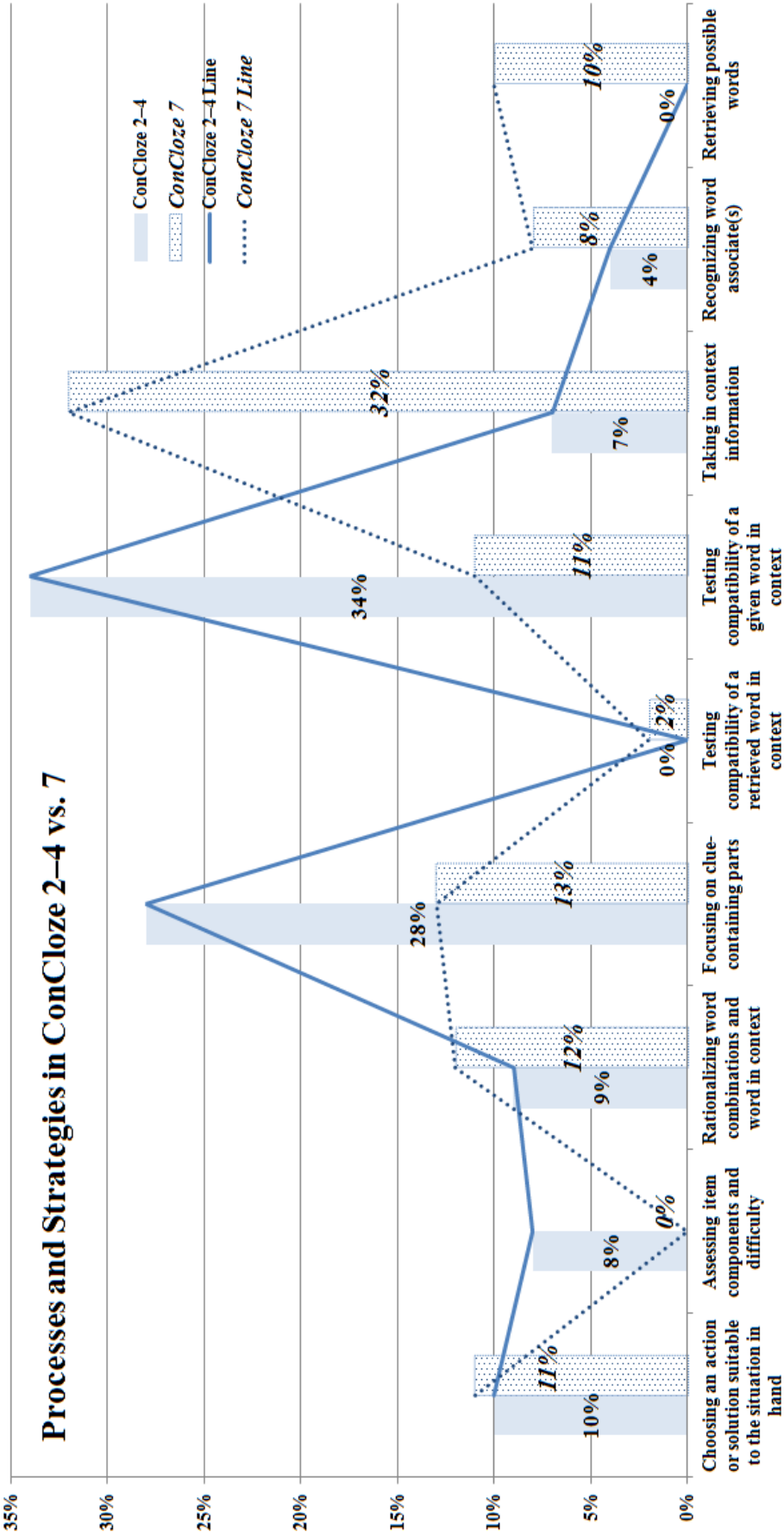


Figure 87 Comparing ConCloze 2–4 and 7 processes and strategies

In Figure 86, two concurrent processes are constructed anew in this inquiry, the first of which is *Retrieving possible words*. This process garners 10% of all the verbalization segments identified (11 of 108). The criterion is when the respondent appears to be deep in thought near the KWIC blank. This is usually followed by the researcher's verbal nudge, their attempt to produce an answer, or their acceptance that they do not know the answer. For example, in Figure 88 below (ComUnit 12), Nina paused at the KWIC blank in Line 6 of Item 2 and then accepted that she did not know the answer. An interpretation is that Nina was likely to have gone into deep processing at the KWIC position, trying to find a right word for it, before recognizing her inability to come up with a probable answer.

Name	Com Unit	Key	Process Level	Process/Strategy	Hint	Verbalization
Maya	11	social	Concurrent	Retrieving possible words	P	[Line 1 again] 'which the categories of Islamic anthropology [E] and [E]' [P] 'the'? [P] 'the' is the word. [Maya wrote the word 'the' in the KWIC blank of Line 1] Something so simple.
Nina	13	social	Concurrent	Retrieving possible words	P	[Line 2 again] 'other sexual and drug ['use' skipped] behavior and' [P] 'ethnic' maybe. [P] No, doesn't make any sense. [Giggling]
Nina	12	traumatic	Concurrent	Retrieving possible words	P	and then [LA 'enduring somewhere?'] hmm. [Line 6] 'was clear that the woman's [P] experience' No, I have no clue. [P] Other again, yeah. It's like we have like the same words all these lines, but I can't find the words that fit in all of them.
Maya	12	social	Concurrent	Testing compatibility of a retrieved word in context	P, E, Inserting words produced	[Line 1 again, with her response constructed 'the' filled on the paper] [LA 'analysis'] are to be sort of operated and investigated'
Maya	13	social	Concurrent	Testing compatibility of a retrieved word in context	P, E, Inserting words produced	[Line 2 again] 'other sexual and drug use behaviors, and' [Maya was going to start writing the word 'the' (constructed) in Line 2] Ah, no. Wouldn't go in, that word. [laughing]

Figure 88 New processes in ConCloze 7

Given that the process *Retrieving possible words* is constructed anew in this inquiry, its relative position of occurrence is also worth considering. All the verbalization segments of this process are found to occur before the option sheet is handed out. An inference is that the respondents might seek to retrieve a plausible word from their mental lexicon whilst having no options provided then yet. Likewise, shown in Figure 86 (page 251) and Figure 87 (page 252), the process does not appear to have occurred in any of the verbal reports back in the selected-response ConCloze 2–4 (cf. also Figure 72 for an analysis of processes and strategies in ConCloze 2–4 items). This seems to indicate that

Retrieving possible words is dependent on engaging in the task in which no option words are yet provided. On these accounts, the process *Retrieving possible words* could be inferred to reflect test-method variance and thus be peripheral to the core score meaning for the ConCloze item type.

The other new process constructed in this inquiry is *Testing compatibility of a retrieved word in context*. This process garners 2% of all the verbalization segments identified (2 of 108). The criterion is when a respondent explicitly tests whether the word that they have come up with fits in a concordance line. In Figure 88 above, the process is observed when Maya came up with the determiner *the* and was seeking to write it in Lines 1–2 of Item 1. However, it is identified only in one verbal report by one respondent; it does not gain sufficient evidence in support of the construct relevance and thus would be argued as test-method variance. That is, a respondent would most likely seek to test a word they retrieve from their mental lexicon when the task requires them to, a situation observed when no option sheet is offered yet. Given the insignificance in size and prevalence, this process is negligible towards the main construct definition.

In addition to the new processes constructed, Figure 86 (page 251) also shows several processes in this inquiry that are similar to those from ConCloze 2–4 (page 230). In defining the core construct, the proportion of these processes are compared in Figure 87 (page 252). It is discovered that the process *Testing compatibility of a given word in context* obtains the highest proportion of verbalization segments (34%) in the selected-response format. By contrast, the process *Taking in context information* garners the highest proportion of occurrences (32%) in the current constructed-response format. Adjacent or remote, the latter process is found in the analysis to always be followed by the processes *Retrieving possible words* and *Testing compatibility of a retrieved/given word in context*. It can thus be argued that the take-in of context information related to the KWIC is a prerequisite for the testing of word–context compatibility. Moreover, it is worth emphasizing that the testing of the compatibility of a given word in the concordance is very likely to be unable to take place meaningfully without processing context information. This means that regardless of the types of expected response, a core process that must be performed in ConCloze would be *Testing a meaningful compatibility of a word in context*—a process that merges the two processes receiving the highest proportion of occurrences in the selected-response and constructed-response formats.

With regard to test-taking strategies, *Focusing on clue-containing parts* receives the second highest proportion of occurrences in Figure 86 (page 251). Applicable to both of the test formats, an inference is that ConCloze examinees seek to allocate the capacity of their working memory and information processing to the parts they believe are significant to task completion. Based on a clue analysis (page 239) and a contrastive analysis (pages 242–245), the parts could be those containing context clues associated in either form or meaning and use with the missing target word. Accordingly, it could be argued that ConCloze engagement involves an on-line selective processing of context clues.

4.3.5 Concluding claims and decision

In Sections 4.3.2–4.3.4, there have been several inferences made based on the evidence from test and item responses. They are summarized in Table 44 in no particular order.

Table 44 Major ConCloze 7 evidence and inferences

Entry	Evidence	Inference
1	Process <i>Testing compatibility of a given word in context</i> prevalent across items from ConCloze 2–4	Active testing of the lexical meaning of an option word given its related information in the concordance
2	Strategy <i>Focusing on clue-containing parts</i> prevalent across items from ConCloze 2–4	Strategic concentration on significant context clues
3	Strategy <i>Choosing an action or solution suitable to the situation in hand</i> taking 10% of verbalization segments	Meta-cognitive planning and decision-making inherent in task engagement
4	Natural collocates in the item prompt and in the respondents' phrasal checks	Collocates as activators of association with the target word in mental lexicon
5	A low performer focusing on syntactic slots	Localized processing upon partial knowledge; grammatical knowledge as marginally construct-relevant
6	A high performer systematically accumulating global context clues	Individual and compositional lexical-semantic content, and collocates as context clues
7	Convenience observed in dealing with the paper-based test	Paper-based ConCloze feasible
8	No correct answer produced for the open-ended format	Constructed-response format in strong form infeasible
9	Process <i>Retrieving possible words</i> taking place only before the option sheet is handed out	Test-method variance
10	Process <i>Testing compatibility of a retrieved word in context</i> occurring in one verbal report	Test-method variance
11	Process <i>Testing compatibility of a given word in context</i> occurring highest in selected-response format, and process <i>Taking in context information</i> in modified constructed-response format	Core process <i>Testing a meaningful compatibility of a word in context</i>
12	Strategy <i>Focusing on clue-containing parts</i> occurring second highest in both of the test formats	On-line strategic processing of selected context clues in ConCloze engagement

In conclusion, the findings for substantive fine-tuning in this collection of construct inquiries can be categorized into three areas of significance. First, the Grounded Theory-oriented processes—constructed anew and reapplied—can account for the verbalizations to all of the multi-variant items from ConCloze 2–4. Most importantly, their proportions applying to the verbal reports are largely comparable across the items (e.g., page 230). Finding processing comparability can be significant because it represents applicability of the test-taking processes across multiple sets of task content – generalizability of the substantive aspect of construct validity.

In addition to the substantive generalizability, a second area of significance lies in a combined effort to enrich an exploratory content analysis of contrastive task

engagements with a top-down identification of word information. It is found that knowledge of individual and compositional lexical content, and knowledge of collocation are likely to be primary language domains tested, with world knowledge and syntactic knowledge peripheral to the core construct proficiency. Being able to synergize a bottom-up analysis of task performance and a top-down linguistic analysis can be significant because the finding can represent fidelity between the two approaches to construct modeling, thereby indicating the applicability of theoretical explanation to substantive empiricism – generalizability of the substantive content to target domains in real language use.

A final area of significance is about constructing Grounded Theory-oriented processes in a new test mode and method. It is discovered that a paper-based, modified constructed-response format of ConCloze is likely to share the majority of test-taking processes and strategies. While a few variations in processes are found in the new format, potentially caused by inherent test-method variances, the core processes and strategies are comparable across the selected-response and modified constructed-response formats. Combining those from the two, testing a meaningful compatibility of a word in context is likely to be a primary process, with an on-line strategic processing of selected context clues. Finding common underlying processes can be significant because this can represent their applicability between test modes and test formats – generalizability of substantive processing across facets of construct measurement.

In Chapter 1, the test purpose is hypothetically set to be a proficiency test on professional and academic English grammatical and vocabulary use (page 23). In light of the summary, the test purpose can be appraised as follows. Up until ConCloze 6, the test purpose has been narrowed down to a proficiency test with a unique combination of knowledge domains in vocabulary use tested. Emphasis is on lexical-semantic knowledge, and knowledge of word association, which may subsume knowledge of collocation and knowledge of synonymy. With the current findings from a series of construct inquiries added up, the importance of the knowledge of collocation is highlighted, and that of the knowledge of synonymy demoted. One new dimension of lexical-semantic knowledge is also identified: knowledge of individual and compositional lexical-semantic content, a domain which may be deemed related to semantic prosody and granularized out of the knowledge of lexical-semantics. In conclusion, the ConCloze item type developed in this study may be argued as a proficiency test on vocabulary use,

particularly in terms of lexical-semantics, individual–compositional lexical-semantic content, word association, and collocation. In this part of substantive fine-tuning, no evidence can be found to uphold the construct centrality of the dimension of grammatical use originally theorized in the hypothetical construct. Rather, knowledge of grammatical structure is categorized as construct-peripheral.

In the rationale (page 226), a need to address inadequacy in substantive evidence is emphasized. In light of the substantive findings, the test-score interpretations could be deemed having an improved power of generalization towards the universe of permissible observations. For example, they have addressed test-mode and test-method bias and thus possess generalizability to more than one test format. Given the generalizability based on multiple validity facets, it can be argued that the inadequacy has been sufficiently addressed and the test-score interpretations are likely to have validity. Considering limited research resources, it is decided that this validity investigation should stop.

Chapter 5 Concluding Remarks

5.1 Summary and Discussion

In this thesis, validity of the score interpretations for a ConCloze test has been investigated. ConCloze is argued in Chapter 1 as a new item type, suggesting that a score meaning that is rigorously defined is of importance (pages 4ff.). In Table 45 below, some of the interpretations are gathered for appraising the aspects of the construct validity dealt with (cf. Kane 1990; 1992 for an all-available approach to constructing a validity argument). For example, in ConCloze 6, WAF scores are analyzed alongside ConCloze scores, addressing five aspects of the validity. First, their scores are found significantly correlating, providing evidence for the external aspect of validity (pages 216f.). Secondly, the correlation means that the domains of knowledge tested in WAF—synonymy and collocation—are also likely applicable to ConCloze, thereby addressing the content aspect of validity. Thirdly, a subsequent investigation into their principal components demonstrates that the responses of the two tests do not load on the components of each other (pages 222f.). Rather than negating the domain applicability, their relation is redefined: knowledge of synonymy and collocation might be tested in ConCloze, but the constructs of ConCloze and WAF could differ in the aspects of proficiency involved. An inference then is that the ConCloze item type could invoke a unique combination of aspects in proficiency in vocabulary use (page 223; to be elaborated later), thereby addressing the substantive-validity aspect. Fourthly, in light of the principal-components analysis, the item responses of ConCloze appear to load on a single component. This thus shows internal consistency among the responses analyzed and therefore forms evidence of the structural aspect of validity. Lastly, considering the results of the correlation analysis and the principal-components analysis, another aspect of construct validity addressed is generalizability—applicability of interpretation across two analyses.

Table 45 ConCloze-score interpretations with validity aspects emphasized

Investigation–Interpretation	Aspect of Construct Validity Addressed				
	Structural	Content	Substantive	External	Generalizability
ConCloze 1: Quantitative Prototyping					
Responses to univariant items are likely governed by a competence involved with knowledge of lexical-semantics	✓	✓			
ConCloze 2–4: Qualitative Prototyping					
Item components processed are comparable across respondents, which can be accounted for by similar test-taking processes			✓		
ConCloze 5: Field-testing					
Influences of demographic variables over the scores are negligible			✓	✓	
Semantic unrelatedness of options and availability of concordance-based clues can distinctly lower item difficulty			✓		
ConCloze 6: Operational Use					
ConCloze 6 can be equated and co-calibrated with ConCloze 5	✓				✓
Influences of demographic variables over the scores from another ConCloze test are also negligible			✓	✓	✓
L1 can be a surrogate for the intensity and extent that the ConCloze competence is acquired			✓	✓	
Associations between English self-ratings and the scores of another ConCloze test are also significant			✓	✓	✓
WAF scores significantly correlate with ConCloze scores		✓		✓	
ConCloze principal component does not load on WAF components	✓		✓	✓	✓
ConCloze 7: Substantive Fine-tuning					
Test-taking processes newly constructed and from ConCloze 2–4 can account for all the 60 verbal reports.	✓		✓		✓
Knowledge of individual and compositional lexical-semantic content and knowledge of collocation are likely parts of the competence tested		✓		✓	✓
Core processes can be produced out of the selected-response and modified constructed-response formats	✓		✓		✓
Count	5	3	9	7	7

It can be concluded from Table 45 that all the aspects of the construct have garnered at least three different investigations based on different samples of respondents and facets of measurement. Given this, the research could likely be claimed to have

sufficiently investigated the construct validity of ConCloze-score interpretations, which will be used later for refining the hypothetical construct set in Chapter 1.

In light of the sufficiency appraisal of validity aspects, the operational research questions from Table 1 (page 16) can be answered in Table 46 below. It is found that all but Questions 3 and 7 have been addressed by at least two test versions, co-constructing the adequacy in dealing with the corresponding operational question. As to Question 3, use of an item prompt and the options was observed in all the verbalizations while the respondents were engaging with the test tasks, and the accuracy is confirmed when the verbal reports are transcribed for data processing. As for Question 7, owing to limited research resources, only one criterion test is administered alongside ConCloze. Yet, two criterion-related inferences are found to reflect the hypothetical construct (page 19) accurately, to the extent that the relationship between ConCloze and WAF can be defined by two analyses without gross construct irrelevance observed. In sum, it could be claimed that the operational research questions have been addressed sufficiently.

Table 46 Evaluating the operational research questions

No.	Question	Aspect
1	Are item responses internally consistent?	<i>Structural</i>
Answer: Yes, e.g., as evidenced by the reliability index in ConCloze 1, and single-factor loading in ConCloze 6		
2	What domain(s) do the task engagements involve?	<i>Content</i>
Answer: Multi-faceted word knowledge, as evidenced by effects of semantic components of the options in ConCloze 5 (lexical semantic knowledge), and systematic phrasal checks with natural collocates in ConCloze 7 (lexicogrammatical knowledge)		
3	What item component(s) do the examinees use in task engagement?	<i>Substantive</i>
Answer: All the language components in the concordance lines of the item prompt and in the options, as observed in 60 verbal reports in ConCloze 2–4		
4	What process(es) do the task engagements involve?	<i>Substantive</i>
Answer: <i>Testing a meaningful compatibility of a word in context</i> , as constructed in a Grounded Theory-oriented fashion based on a merger between the core process in the selected-response format in ConCloze 2–4 and the counterpart in the modified constructed-response format in ConCloze 7		
5	Is item difficulty affected by variation in task content?	<i>Substantive</i>
Answer: Yes, as evidenced by a reduction in item difficulty when semantically unrelated options are used in ConCloze 5, and a dispersal of IFs by the gradient-frequency test design in ConCloze 6		
6	Is there consistency in item responses and processes across occasions?	<i>Generalizability</i>
Answer: Yes, as evidenced by crossing regression lines accounting for ConCloze 5 and 6 items, and shared processes underlying the task engagement in all of ConCloze 2–4 items		
7	Are ConCloze scores significantly associated with Read's (1998) WAF scores?	<i>External</i>
Answer: Yes and no, both in ConCloze 6. Yes, as evidenced by a significantly high correlation coefficient. No, as evidenced by no factor cross-loading between the two tests.		

The following is a summary of the score meaning, which will be elaborated later in light of an appraisal of the hypothetical construct. In Chapter 2, the construct competence is theorized to be multicomponential (page 35). From the validity investigation, the construct domain is found to be very likely composite: at least lexical-semantic knowledge, knowledge of synonymy, knowledge of collocation, knowledge of grammatical structure, world knowledge, knowledge of word association, knowledge of semantic prosody, and knowledge of individual and compositional lexical-semantic content are likely to be involved in test-task engagement. Albeit with unequal weight and hierarchy, the competence domains are generally operated integratively, varying in accordance with test-task content and the examinees' differing levels of proficiency.

Considering the construct domains summarized, the test purpose from Chapter 1 can be meaningfully refined as follows. The purpose has been hypothetically set to be a proficiency test on professional and academic English grammatical and vocabulary use (page 23). In this validity investigation, the test purpose has been narrowed down to a proficiency test with a unique combination of knowledge domains in receptive vocabulary use tested. Organized in Figure 89 below, two areas that are found to be consistently mobilized and accordingly central to the construct proficiency are lexical-semantic knowledge, and knowledge of word association, which will be discussed in turn.

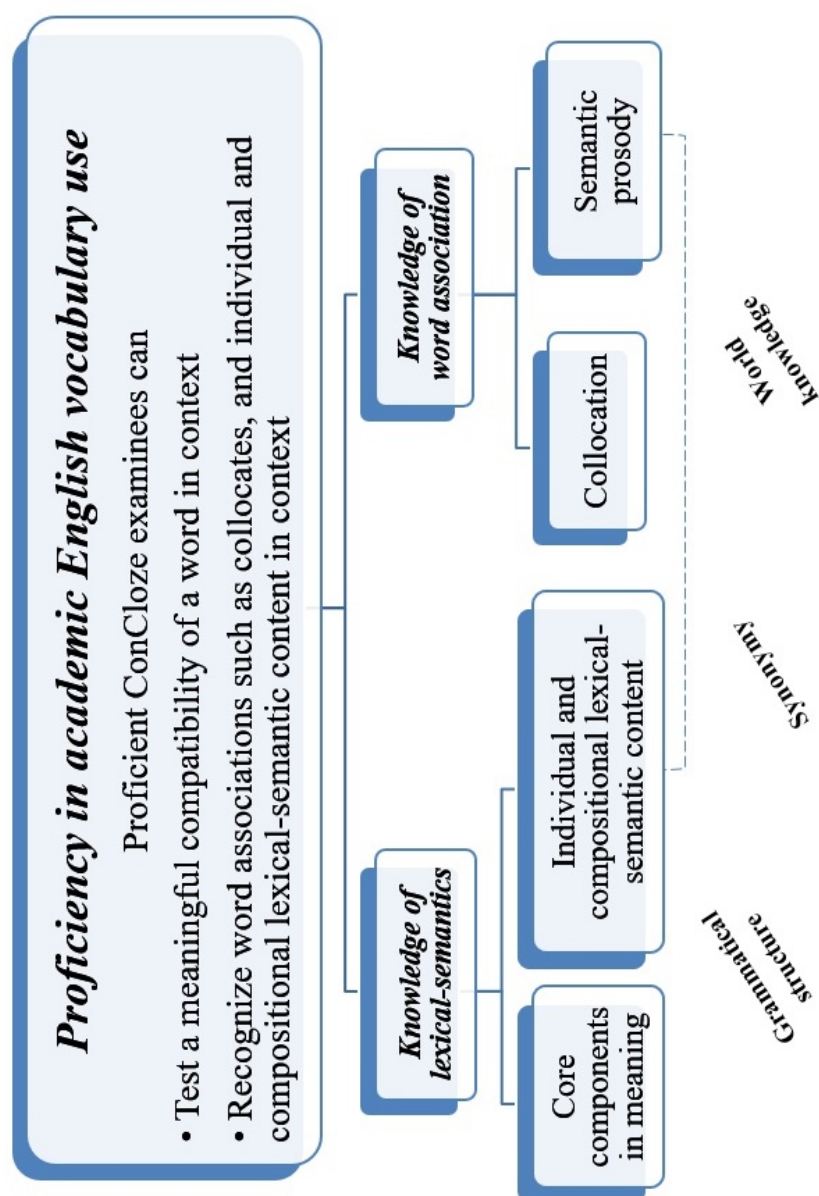


Figure 89 Potential processes and domains of knowledge identified for the proficiency testing

With regard to knowledge of lexical-semantics, knowing individual and compositional lexical-semantic content is found to be one of its most fundamental subordinate domains. Displayed in Figure 89, mobilizing the knowledge is when content of words is processed individually and in combination – an aspect of linguistic processing which examinees need to have in task engagement. This can be exemplified by two perspectives: bottom-up and top-down. Regarding the bottom-up perspective, words in a concordance prompt are found to be read both in full and in part, often with repetitions (e.g., page 138). This suggests, again, that the lexical-semantic content of words is likely to be processed during active engagement of the tasks (page 245). Regarding the top-down perspective, for example, the target word in Section 4.3.3 (page 238) is *recreational*, and one of its concordance-based descriptive texts is *capture their interest*. The descriptive text is interpreted such that it might serve as a clue to the word, suggesting their synergetic processing that goes beyond a single-word level. Limited or extensive, knowledge of lexical-semantic content may also be enhanced by world knowledge, which is found to be construct-peripheral considering its possibly optional role in task processing. Because vocabulary size is generally a function of general English proficiency (e.g., Laufer & Nation 1999: 38), it may thus follow that knowing lexical-semantic content of more words is likely to contribute to a higher proficiency level of academic vocabulary use than having the knowledge of fewer words.

Apart from the lexical-semantic content, lexical-semantic knowledge is also found to expand beyond making sense out of a single word and a string of words. It can also go into relations between words, which could be conspicuous in the selected-response format. The discussion may begin with two cases of the semantic components of option words, the first of which is when some of the option words are synonymous or share many of their semantic components (as in IV2–6 in ConCloze 5, e.g., pages 179ff.). Those core components in meaning are found to likely be vigorously processed in light of the clues as they can be decisive in distinguishing one option word from the others (e.g., page 179). The other case of semantic components of option words is when the words are semantically distant or unrelated (as in IV1 of ConCloze 5, pages 179ff.). Completing those tasks appears to be much easier amidst word information provided in the prompt.

Considering the two cases of semantic relations among option words, two domains seem to be involved in test-task engagement. First, given the consistency in processing semantic relations in light of the concordance-based word information,

knowledge of core components in meaning seems central to the construct proficiency. The knowledge is thus presented in Figure 89 (page 264) as the other subordinate domain of lexical-semantic knowledge. On this account, also illustrated in Figure 89, the core process for the item type is to test a meaningful compatibility of a word in context, which is likely to be required for receptive proficiency in vocabulary use.

A second domain in light of considering semantic relations among option words is knowledge of synonymy, which may or may not be called for, depending on the meaning content of option words. As contrasted when options are not semantically related, the lack of consistent mobilization seems to indicate that synonymy knowledge is not always vigorously mobilized. While its consistency could be observed across test items as in ConCloze 1 (pages 97f.), such observations may be attributed to variation in test-task content, particularly in terms of option design. Based on this optionality to the construct proficiency, knowledge of synonymy is not central to the construct and thus organized in Figure 89 (page 264) as construct-peripheral.

In light of making sense out of semantic relations among option words, an observation from the modified constructed-response format (Section 4.3.4) is also worth recapping. It is observed that before the option sheets were handed out, the respondents attempted to produce a possible answer in separate items by forming speculations out of the content of the concordance prompt (page 250). The semantic aspects related and shared among the lines seem to be used for forming such speculations. Accordingly, knowledge of core components in meaning is likely to also be involved in the construct proficiency even in a modified constructed-response format. This thus implies that while knowledge of synonymy is invoked primarily because of option words and hence is construct-peripheral, knowledge of core components in meaning may be deemed intertwined with the conceptual representation of a target word as derived from the prompt (cf. page 241 for a visualization of the prompt–target word association). Accordingly, knowing core components in meaning has a wide applicability towards matching with the concordance-based word information and thus central to the construct proficiency.

In addition to knowledge of lexical-semantics, knowledge of word association is also found to be tested in the ConCloze item type. Illustrated in Figure 89 (page 264), a consistently mobilized subordinate domain of word association is knowledge of

collocation. The findings indirectly indicate an important role of the knowledge of collocation in forming a lexical network so as to give a hint about the target word (cf. page 214 for a briefing on this topic). For example, particular phrases and partial concordance lines containing collocates are picked up systematically, albeit unknowingly, by the examinees (e.g., page 138). This allows for a construction of the process *Recognizing word associate(s)*, and the strategy *Focusing on clue-containing parts*, for instance. Relating to knowledge of collocation, such processes are identified in both qualitative parts (a single-item analysis of verbal reports in ConCloze 2–4, and a multiple-item analysis of verbal reports and a corpus-informed substantive contrast of a respondent sample in ConCloze 7) and quantitative parts (an inference based on difficulty-lowering effects of semantically distant option words in ConCloze 5, and a correlation analysis of WAF–ConCloze scores in ConCloze 6). On these accounts, knowledge of word association, especially its subordinate domain of collocation, is likely to be a construct-relevant domain for testing receptive proficiency in vocabulary use.

Apart from the knowledge of collocation, another domain of knowledge subsumed under knowledge of word association is semantic prosody. Reviewed in Chapter 2, this is when a meaning co-occurs frequently with a word. In Figure 89 (page 264), the domain is linked to knowledge of individual and compositional lexical-semantic content because both individual words and their combinations could each construct a meaning that may be expected in frequent use with a particular target word and thus can be counted as part of word association.

In conclusion, the test purpose may be refined into a receptive proficiency test on academic vocabulary use, testing lexical-semantics, and word association. While knowledge of grammatical structure is found to likely be mobilized by one respondent with a low proficiency level, it is not systematically mobilized by the other respondent with a higher proficiency level in the same inquiry (Section 4.3.3). For this reason, knowledge of grammatical structure is categorized as another construct-peripheral domain of knowledge in Figure 89 (page 264). This thus means that in this study, no evidence can be found to uphold the construct centrality of the dimension of grammatical use originally theorized in the hypothetical construct.

In addition to appraising the adequacy of the score meaning defined as well as how it could help in refining the hypothetical test construct, appropriateness of inferences

is also considered in this research. Another pillar of test validity, it is intended to support the plausibility of the current validity argument (cf. Messick 1993 for inseparable pillars of construct validity). In the test-development process, the appropriateness of primary actions and evidence has been appraised in Table 33 (page 183). For appropriateness of the thesis-level inferences, key areas of research design and data analysis are summarized in Table 47 below, in which two questions are posed. The first is whether the inferences can represent particular facets of the corresponding construct measurement. For example, several types of inquiry such as score descriptions and analyzing the content in verbal reports are used for interpreting item responses and other observable behaviors. The common ground of these inquiry types is to understand the nature of the construct. Using several types of inquiry means that the total interpretation would not be too narrow and can stand out amidst inherent test-method variances. Based on the areas summarized, an appraisal is that all the construct-related facets are represented by more than one component. In light of this heterogeneous representation, it may be claimed that the sampled test performance and task processing could represent the construct proficiency appropriately.

Table 47 Cases for appropriateness of score interpretations

Representation of Construct-related Facets	Relevance of Construct Interpretations
Adequacy check: Two investigations for the same facet	Examinee sample: Population relevance
Format: Selected-response, and modified constructed-response	Construct definition: Spec-driven evidence-centeredness
Delivery mode: Internet-based, and paper-based	Response validity: Testing usability, adjusting item-difficulty for intended population
Sampling: Convenience and snowball with mixed L1 groups	Nomothetic empiricism: Framed with English self-ratings
Sampling adequacy: Iterations with different samples sufficient for analyses required	Nomothetic empiricism: Negating non-language-related variables such as age, education level
Word sample: Narrow frequency bandwidth, and wide-range frequency levels with interspersed omission	Nomothetic empiricism: Fine-tuned with face-validity corpus insights and contrastive high–low performers
Type of inquiry: Descriptive, inferential, log-linear, textual, content, observatory, and semi-structured interview	Nomothetic empiricism: Underlying processes in task performance
Inquiry perspective: Bottom-up generalizability, and top-down corpus-based empiricism	Nomothetic empiricism: Relating to criterion test WAF

The second question asked in Table 47 above is whether the interpretations are relevant to the construct. For example, the respondents are screened for their population relevance, so that their task performance could be argued as relevant and likely to be generalizable to that of the examinees in the target population. Another example is when WAF scores are used for redefining relations with ConCloze scores, in which the ConCloze construct is argued as part of the general proficiency in vocabulary use that encompasses much more than knowledge of synonymy and collocation (pages 222ff.). The redefinition means that the domains of synonymy and collocational knowledge, which have been argued as applicable across the two tests, become part of the composite score meaning (page 224). Notwithstanding, when a maximal contrast of substantive processing is performed in ConCloze 7 (pages 235ff.), the knowledge of synonymy is identified as unlikely to be central to the construct and hence demoted to be only construct-peripheral.

In light of the multi-faceted investigations into construct definition, an appraisal can be that the construct-irrelevant variances have been singled out of the construct-related descriptions. Accordingly, it may be claimed that the score inferences are appropriately relevant to the construct proficiency tested in the ConCloze item type. In using multi-faceted mixed methods, strategic links between measurement facets are created based on the items that the test versions have in common. Visualized in Figure 90 below, the strategic links allow synergizing quantitative and qualitative strengths in exploring the ConCloze trait. In sum, the construct-relevant performance for the receptive proficiency in academic vocabulary use has been sampled and likely to be represented appropriately in the score interpretations of this study.

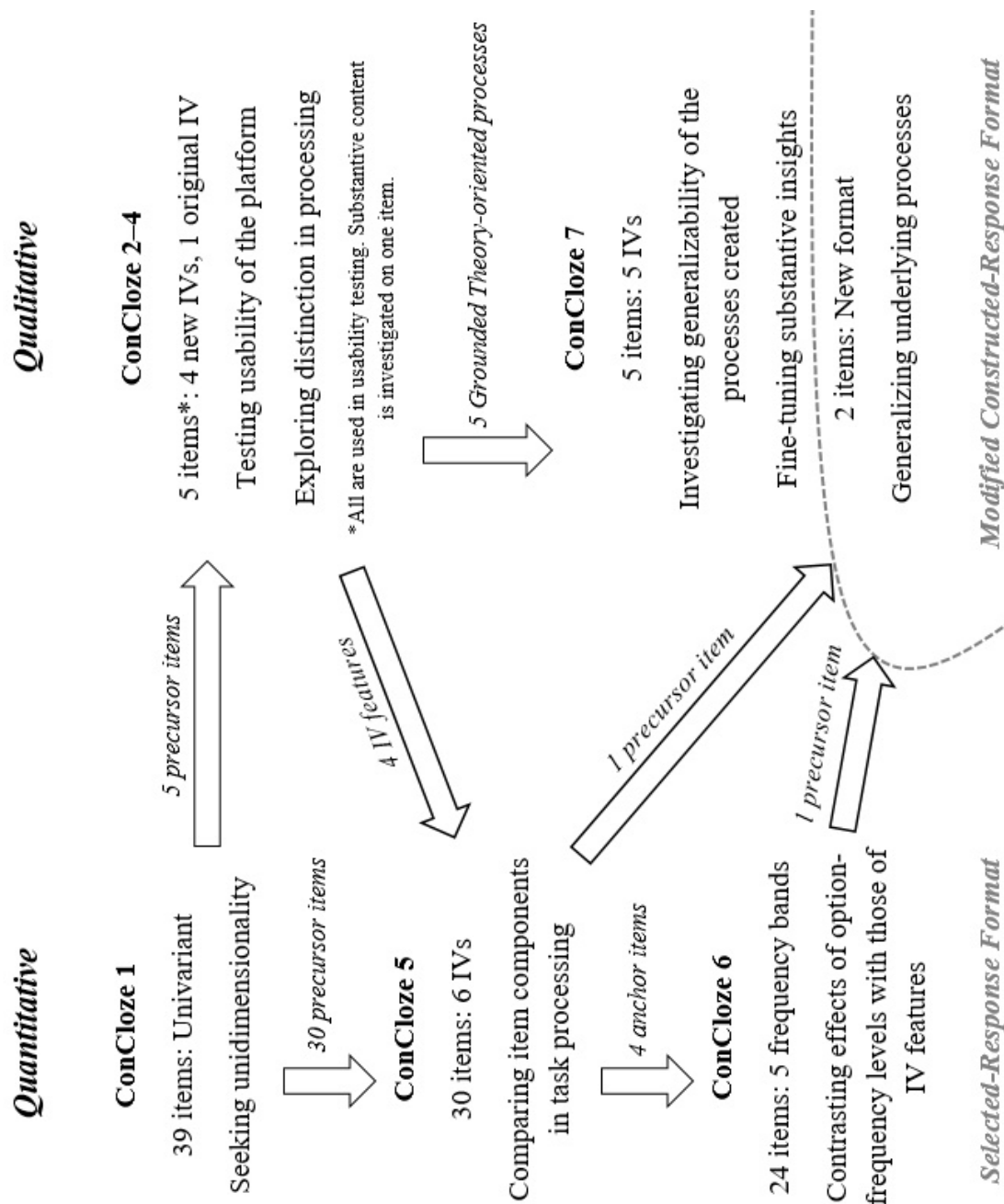


Figure 90 Spec-driven substantive investigation

Two key elements of test validity are adequacy and appropriateness of test-score interpretations (cf. page 2). Considering the adequacy and appropriateness appraised thus far, it may be claimed that the ConCloze-score interpretations have validity. Consequently, the gap in the literature that has been identified in Chapter 2 (pages 31ff.) may also be argued as appropriately addressed by this validity investigation.

Given the validity of test-score interpretations, the theoretical framework in Chapter 2 (Section 2.3 Framework for Construct Measurement, and Section 2.4 Lexical Priming in Cloze Procedure) could also be revisited. The aim is to advance a construct model for this item type based on both conceptual speculation and construct empiricism. Generating a construct model can be useful because it gives meaningfulness to the validity investigation – a generalized and succinct idea of what the scores can represent (cf. Messick 1993 for meaningfulness as a pillar in the very notion of validity; Ross 2012: 223f. for contexts of task performance as part of the validity argument). With the theoretical–empirical fidelity taken into account, Figure 91 below demonstrates the components of the construct proficiency, the core process of which is *Testing a meaningful compatibility of a word in context* (page 255). It is worth highlighting that some of the construct-related aspects are similar to those already presented in appraising the hypothetical construct in light of the test purpose (pages 264ff.). A distinction between appraisal for the test purpose and appraisal for a construct model is that the former appraisal focuses on the usefulness or utility of the test, which is to serve as a receptive proficiency test in academic English vocabulary use. By contrast, advancing a construct model focuses on the meaningfulness of what the scores could be deemed to represent – a take-away message of wide applicability for test validators and language practitioners (cf. pages 2f. and 24ff. for the grounds in support of investigating the construct validity of test-score interpretations for a new item type like ConCloze). Hence, to avoid redundancy, only new information from Figure 91 will be discussed.

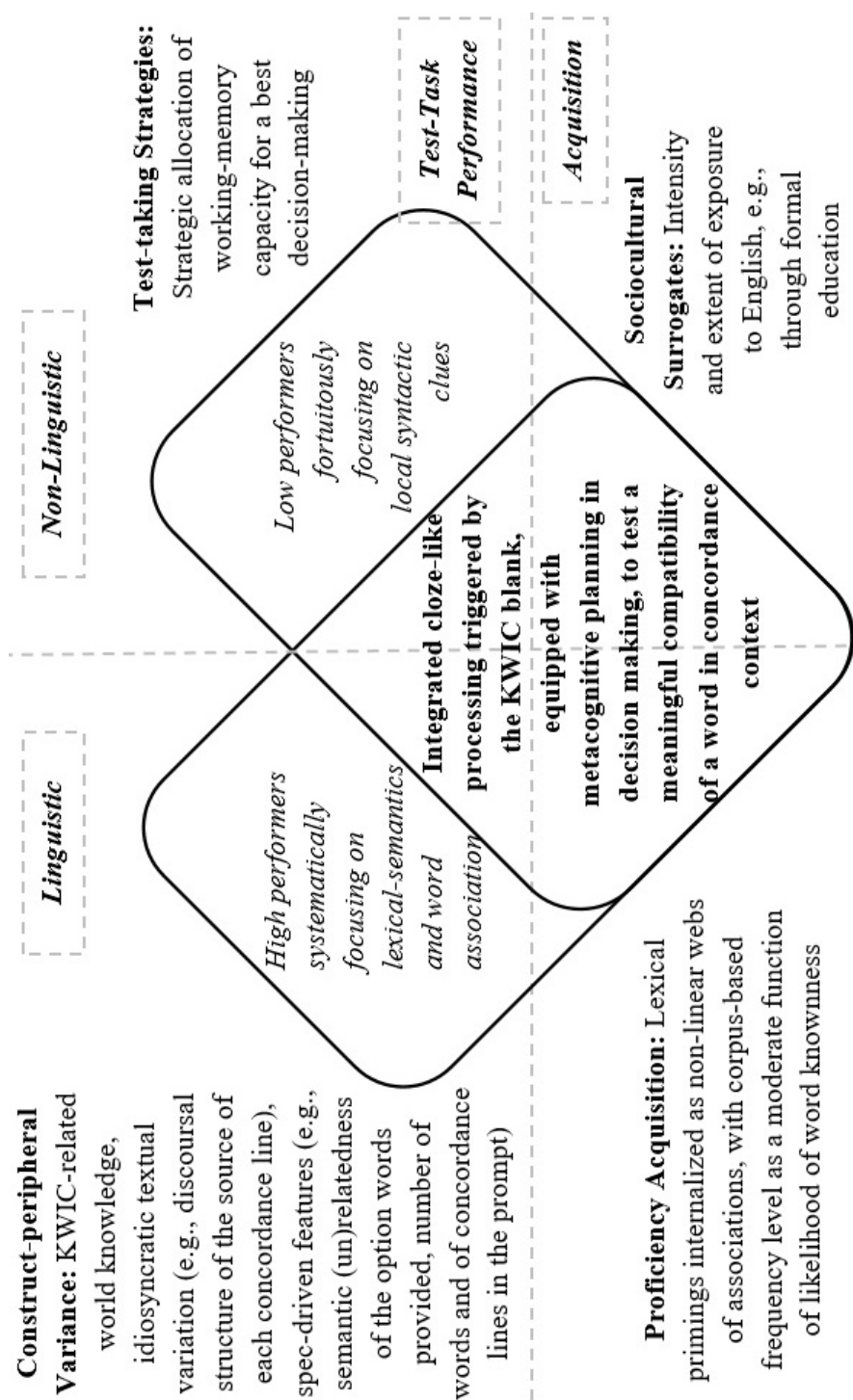


Figure 91 A construct of the ConCloze item type

Reviewed in Chapter 2 (page 54), the cloze procedure is also incorporated into Figure 91 so as to reflect the composite nature of the score meaning (cf. page 224). For example, it could be either knowledge of lexical-semantic content or knowledge of collocation or neither of them but knowledge of semantic prosody and lexical-semantic

knowledge of core components in meaning that are mobilized meaningfully and integrated by individual examinees (cf. also the argument of different aspects of grammatical and vocabulary use best suiting the individual examinees, page 56). A case pertinent to this composite variation is Igor's task engagement, which explicitly involves considering grammatical aspects of the KWIC blank (pages 244 and 245f.). By contrast, proficient respondent Halle seemed to focus on the collocational aspect as well as the lexical-semantic content of the prompt texts provided (pages 242f.). On these accounts, the aggregate effects of context clues towards mobilizing domains of knowledge could be postulated as the driving force for testing the meaningful compatibility for the construct proficiency in vocabulary use.

In acquiring the proficiency tested, lexical priming from Chapter 2 (pages 59ff.) is also incorporated in the linguistic zone of the model in Figure 91. Repeated encounters and productions are argued as accounting for how a target word becomes internalized and its associations such as collocates and semantic prosody are primed for habitually co-occurring with it. A side discovery regarding the priming is that corpus-based frequency levels may not reflect how the words are stored in and retrieved from the mental lexicon (e.g., page 95). During ConCloze engagement, the storage could be activated selectively for word processing or retrieval depending on the concordance-based clues (page 216).

The significance of incorporating the cloze procedure and lexical priming to the construct model is twofold. First, a cloze test usually refers to a text with multiple blanks interspersed, each requiring one word (as in Figure 15, page 54). By contrast, ConCloze features multiple blanks requiring the same word. The fidelity between the theoretical framework and the finding indicates that an integrated processing in the cloze procedure is also likely to underpin the testing of proficiency in academic vocabulary use by ConCloze (pages 55ff.). Thus, the first significance lies in a new context of applicability for the cloze procedure—a meaningful closure of multiple gaps, which has been proposed in Chapter 2 (page 57). Secondly, when taking lexical priming into account, it may be theorized that the priming of a word is not a one-on-one imprint onto the mental lexicon. Rather, the construct definition indicates that one word may be stored with multiple primings of language components (e.g., collocation, semantic prosody, genre). In turn, these primings are apt to be reactivated by stimuli such as context clues in the ConCloze item prompt. As proposed in Chapter 2 (pages 57f.), language testing is thus a new area of applicability of the theory of lexical priming, rather than merely language learning and

acquisition (as in Hoey 2005: Chapter 1). Accordingly, based on this construct model, associations produced by the primings are the determinators of processing normal contextual constraints during task engagement.

The next component of the construct model is when high and low performers' task engagements are contrasted (Section 4.3.3, pages 235ff.). The task processing of the high performer is found to systematically involve word-association knowledge (pages 242f.). By contrast, a low performer may lack a full command of the knowledge and compensate for the lack by focusing on local clues such as syntactic slots (pages 244f.). This contrast in task engagement seems to accord with the theoretical framework in Chapter 2 (page 60), in which a high performer would have a more densely interwoven mental lexicon and thus more nets of primings than a low performer. This would then enable them to effectively link between a target word and its associations as well as related core components in meaning that are described in the concordance-based task content.

The last linguistic component of the construct is a function of inter-item variance: the domains such as world knowledge that is related to the KWIC may also be mobilized in ConCloze processing. They are included in the model as construct-peripheral variance—ancillary but not entirely indispensable—because they are not found to be systematically invoked for task completion (e.g., Section 3.3.4 Substantive content, page 124, and pages 239f.).

When the examinees deal with the tasks, they also tend to strategically allocate their processing capacity to selected parts of the concordance which they believe could be significant for relating to the missing KWIC. On-line and unplanned, this allocation is for a best decision-making and may thus be deemed non-language-related in the construct model. The other non-language-related component of the model encompasses sociocultural factors. For example, education level and the status of English in the examinees' countries are posited as surrogates for the intensity and extent of English exposure prior to task engagement (page 204). Albeit non-linguistic, the factors may help to explain the likelihood of knowing words and internalization of their primings.

The significance of the two non-linguistic components of the construct model lies with the fidelity between the theoretical framework and the findings in this validity investigation. In Chapter 2 (page 60), the genesis of the lexical primings is said to be repeated encounters with the words in their context of use. Likewise, the coordination of

domains of competence in the cloze procedure is brought about by extralinguistic factors such as experience, education, and acculturation, in which the co-textual normality is learned (page 55). In ConCloze, education level and sociocultural factors are also found to be likely surrogates of the exposure to words as used interconnectedly in context. Based on the construct model, the proficiency in vocabulary use would thus need word encounters over time and with intensity as in through formal education—the qualitative non-linguistic aspects known for assisting in acquiring multi-word expressions (Adolphs & Durow 2004; Schmitt 2010: 26ff.). Moreover, it is worth stating that strategic competence can be deemed encompassing lexicogrammatical knowledge and sociolinguistic knowledge (Bachman 1990: 84ff.; 1991). Hence, incorporating the strategic allocation of the processing capacity in the construct model can reflect this non-linguistic dimension of the ConCloze construct proficiency.

5.2 Research Limitations

Constructs relating to language proficiency are difficult to define precisely. This is so because they can be context-dependent and usually involve a complicated combination of language abilities (Chapelle *et al.* 2010: 4). Given this, the current study needs to define a broad construct for ConCloze so as not to be context-bound and lose the power to explain task performance across different situations. In doing so, the fidelity in extrapolating to the target domain of competence has been traded off for the power of generalization of the score interpretations. Thus, the construct theory advanced in this study (pages 272f.), albeit aiming for generalizing across facets of construct measurement (e.g., pages 177f.), may fail to fully account for performance in individual test tasks in the future. Further, the construct interpretations aim to be generalizable to varied proportions of related aspects and domains of receptive proficiency in academic vocabulary use. As such, they could not be tailored so as to readily fit a particular test program or curriculum in detail (page 23). On this account, the first limitation is a limited power of extrapolation to the target proficiency as well as a limited utility in fitting test programs outside of this research context.

Apart from trading off the power of extrapolation of test interpretations to the construct proficiency, another limitation is about the approach to inquiring into its nature. In Section 1.3.1, it has been introduced that a pragmatic solution to making score

interpretations for an innovative item type like ConCloze ought to be set as exploratory (pages 16f.). One of the limitations of this approach is that the validity studies can appear less structured, at times slightly chaotic. For example, because there is no precedent of a validated construct for this research to follow, a careful approach to construct definition has been planned using iteration for inductive argumentation (page 66). This is to the extent that, considering the innovativeness of the item type, the very first test version chiefly asks if the item format would ever work in assessing a distinct construct – an inquiry question that would not have been posed in a confirmatory study due to its oversimplicity (cf. Figure 90 (page 270) for a summary of the purposes of the test versions). On the one hand, the iteration has indeed enabled the construct exploration not to start with too narrow a domain of competence, but to sharpen the construct definition along the way (cf. page 66 for a sharper definition intended). Yet, this also requires a systematic yet persistent appraising of the hypothetical test construct in light of the empirical findings throughout the testing process (i.e., in the concluding sections of all the test versions). Eventually, it is thus a series of such persistent appraisals of the construct accuracy that leads to demoting out of the core construct the knowledge of grammatical structure and knowledge of synonymy in ConCloze 7 (page 258).

Besides the chaotic organization of ideas resulted from the iterative design, another example of inherent shortcoming stems from limited resources. Because research resources have to be taken into consideration on a long-term basis for the entire project, the development process as well as the operational test use and application involve only moderate sample sizes for their corresponding purposes (e.g., 13 test-completers in prototyping a newly developed format and giving initial validity evidence in ConCloze 1). This means that only types of analysis needed can be performed, and additional *ad hoc* lines of inquiry are limited, some even requiring future studies to address instead. Moreover, because the sample sizes are not amply vast, there is always a chance of false positives and false negatives merely because the margin of safety is not wide enough to allow for random error variance to run free to a great extent. This thus signifies a fundamental shortcoming: even though with a multitude of consistency aspects observed throughout the study, there can be no guarantee, for example, that the construct domains identified during validity investigation and summarized in Figure 89 (page 264) would be exhaustive.

In addition to a lack of rigid structure for defining the construct, another aspect of inquiry-induced limitation also originates from the test design. In conducting an inductive–argumentative inquiry, several topics of investigation start from those initially identified in ConCloze 1 before expanding in later test versions. This means that the subsequent topics could be limited by the very nature of the quantitative and qualitative prototyping versions and their samples of respondents. On this account, the second limitation is not only a lesser extent of structuredness of the validity argument when compared with a confirmatory study, but also an interdependence between test versions that makes the scope increasingly specific but narrow.

Given an aim to explore an item type little known, it seems reasonable to start the construct inquiry from a broad perspective in ConCloze 1 to a narrower scope in ConCloze 7. This line of inquiry aims to gradually narrow down the range of possible domains of construct proficiency, which may not be methodologically sophisticated. For example, the inquiry may fail to come up with an exhaustive set of underlying factors upfront, for which a controlled experiment could have been conducted. Being unable to control related factors for test-task performance would then mean a limited predictability of actual contexts for task completion (i.e., limited extrapolability of all the variables involved). In short, because of limited research resources, some good ideas that emerge during validation might not be able to be pursued extensively. Not all potential entangled aspects of multicomponential proficiency in ConCloze could be studied either – a seemingly inherent pitfall of a psychometric scale of language proficiency (Nunan 1991). On this account, in addition to an increasingly narrower scope of construct exploration, a third limitation is that the exploratory approach used cannot address such complex methodological issues that could otherwise have been dealt with by an experiment.

Similar to the inability to experiment all the potentially related factors of the construct proficiency, the samples of respondents are likewise obtained with an inherent limitation. Sought through convenience and snowball samplings (page 85), the examinees could be considered self-selecting: those who take the test have chosen to participate in the first place. Generally, such self-selection might be deemed to draw primarily the examinees who are highly motivated, and the test responses observed might not be able to represent as wide a range of proficiency levels as possible.

Considering the potential of self-selected respondents, several measures in test engineering and logistics have been taken. First, test difficulty is gradually reduced such that (a) it becomes appropriately aligned with the average levels of the test completers' proficiency in later versions, and so (b) examinees with a low proficiency level may not feel discouraged to participate in the testing. Another measure is iterative design for multiple test versions across multiple time frames (as in the summary on page 270). This enhances a chance of getting participants of diverse proficiency levels on the whole. Moreover, anonymity and confidentiality are pledged explicitly in the test invitation and on the welcome page of the test, encouraging examinees of both low proficiency and high proficiency alike to feel at ease in participating in the testing.

In addition to the design-based and logistic measures, analyses of item responses are also carried out assessing the possibility of obtaining only self-selected respondents. A first example is concurrent calibration (pages 209f.), which may be deemed to increase the range of proficiency levels for a representative modeling by means of larger samples of examinees. Another example is to check the range of the test performance among test completers across the quantitative versions, which seems to be reasonably wide throughout (scores of 10–28 out of 39 in ConCloze 1 (page 89), 5–30 out of 30 in ConCloze 5 (page 159), and 4–24 out of 24 in ConCloze 6 (page 197)). This wide range potentially suggests (a) some effectiveness of the test delivery in diversifying the prospective proficiency groups, and thus (b) sample representativeness. An implication is that the norms computed from the scores are unlikely to be tightly clustered around the high-proficiency groups or highly motivated ones only.

Notwithstanding the measures exemplified, there can be no guarantee that some of the respondents who dropped out of the testing might in fact be those who had a low level of proficiency or were poorly motivated. This study adheres to BERA's (2011) guidelines for participant dropout (page 86), so it is impracticable to ask the respondents not to leave the testing. Moreover, this study chiefly uses norm-referencing for the reasons of practicality discussed in the framework for construct measurement (pages 40ff.). One of the biggest drawbacks of the paradigm is the influence of the sample characteristics on the scores and therefore test interpretations. For example, if the respondents in one test version all happened to be weak in the construct proficiency, then the test would have appeared to be unrealistically difficult. At the opposite extreme, if all the test takers were those fully proficient in the construct tested, then a delusion would have been that the test

were very easy. Accordingly, it is acknowledged that the ability groups of examinees in this study may not be fully exhaustive for the actual population and its motivation groups, which would include those devoid of the construct proficiency. On these accounts, a fourth limitation is that the samples used would not be able to fully represent the widest range of the population's ability and motivation levels existing in the non-native English population across the globe (cf. also page 85 for an operational definition of the population intended). Even though the chance might not be strong considering all the examples of the measures taken, the possibility of generalizability being compromised to some extent cannot be entirely dismissed.

Similar to the norm-referenced paradigm for proficiency estimation is the view of items for measuring the construct proficiency. In Chapter 2 (page 44), selecting target words in ConCloze tasks is then deemed a medium for estimating the hypothetical construct proficiency in grammatical and vocabulary use. This is in the sense that recognizing the clues in the concordance prompt as belonging to one specific option word would lead to selecting it instead of the others. From the item view, this means that the target words and their concordances in each test version serve as a sample of all the possible situations that the respondents can potentially engage with. One of the biggest drawbacks of this approach is that the items in the test co-construct the norm for estimating their own properties, which would thus suffer bias in sampling to some extent. On the one hand, measures to counter such bias have been taken. For example, to avoid irrationality in pure random sampling, a middle-frequency bandwidth is focused on in ConCloze 1 on the ground of potential usefulness to discriminating language learners (page 73). Later in ConCloze 6, multiple frequency bands with interspersed intervals are used for sampling option words across AVL on the ground of testing effects of frequency level (page 188). Yet, applying principled bases does not guarantee that random error variance, for instance, would not plague the test-score interpretations derived from the norm-referenced estimation of item properties. Moreover, because combinations of a set of target words (Gardner & Davies's (2014) AVL) amidst their contexts can be infinite, a possibility exists in that even with all the items generated in this study combined, item properties as well as subsequent test-score interpretations would be subject to being revisited in light of new items generated in the future. On these accounts, the fifth limitation is that despite the measures of usefulness taken, the possibility of

generalizability of the score interpretations being compromised to some extent, particularly by random error variance, cannot be completely ruled out.

Considering that the respondent samples used would be utterly unable to reflect all proficiency levels and motivation groups possessed by the population, a customized test administration might have been organized for the purpose. For example, a high-stakes ConCloze could have been manipulated, as in a final examination of a university course. The reason is that in high-stakes settings, examinees of all ability levels in the sampled group would feel compelled to complete the test tasks to the best of their ability, no matter what level of general proficiency and language motivation they have. However, it would seem to be reckless to administer a high-stakes test whose score meaning was yet undetermined in the course of validity investigation. For example, it seems unfair to the students enrolled in the program that their summative assessment would rely on a test format that they have never seen, let alone if it would be able to test the language content as specified in their program syllabus. On this account, also related to the previous limitations, a sixth limitation is that a high-stakes ConCloze test could not be administered for samples of respondents representing all groups of language ability and motivation in the population.

In addition to the limitation pertaining to high-stakes testing, another limitation is on implementing undercover techniques for fully representative respondents. Using undercover techniques can be divided into two following ways. First, ConCloze testing is low-stakes to the respondents in this study (page 92). A related undercover technique would be to administer the test as a purportedly high-stakes one when in fact it is not. However, using such a technique may breach research ethics in terms of deliberate false information and harmful effects such as excessive test preparation and stress unnecessarily caused to the respondents. The second undercover technique considered is to mix the ConCloze item format in the same battery as other established formats such as reading passages with multiple-choice questions. However, the uniqueness and innovativeness of the item type might become a weakness instead: the very format of ConCloze would make itself distinctive and stand out when in use amidst other established item types.

In light of undercover techniques, another similar idea is to administer ConCloze inside established test programs such as TOEFL. In the same test administration, the test

takers would be informed of experimental or research questions as part of the test battery. Even with other filler items used alongside (those aiming to disguise the real item being focused on in a particular study), the innovative use of concordances could cause suspicion among test takers. This would then result in variations in item responses due to their recognition of ConCloze being the part giving them no effect on their scores. On these accounts, the seventh limitation is the inability to tacitly implement undercover techniques for a largest possible range of respondents' proficiency levels.

Apart from the limitations on ability range and participant motivation, another limitation is on the online test platform SurveyMonkey (cf., e.g., pages 85 and 110 for discussion related to the platform). In view of enhanced generalizability, the platform is used for quantitative versions (ConCloze 1, 5, and 6), intended to garner responses from examinees of different first-language backgrounds in multiple countries. Another usefulness anticipated is that task engagement is not time-limited. This (a) allows the respondents to fully and presumably separately mobilize their construct proficiency at their own pace and (b) mitigates construct-irrelevant variance caused by, for example, haste or a necessity to manage time. A major problem inherent in these practices is that there is no proctor overseeing their online engagement. Some of the examinees may, for example, use dictionaries, ask friends, and search the internet for information. As far as an anecdote of my personal communication with the SurveyMonkey provider is concerned, simultaneous web browsing, for instance, cannot be restricted. This might be because such control would require a root access of the web browser and the operation system, an undue compromise of digital safety on the respondent's part. On this account, the possibility of an eighth limitation exists: some of the item responses could be plagued with invalid task performance.

Given the possibility of response invalidity, several measures have been taken in order to augment the effectiveness of eliciting responses that are legitimate. For example, in designing research iteration, several items are reused and/or modified, which serve as strategic links between test versions (cf. Figure 90 (page 270) for some spec-driven links for substantive investigation, particularly between the quantitative and qualitative versions). Hence, the validation process as a whole has obtained both quantitative and qualitative responses to multiple items that are comparable. It is also worth restating that each qualitative item response involves verbalizations during a one-on-one elicitation, which is proctored and excludes help from reference materials. That is to say, the patterns

of responses to the unproctored quantitative versions are complemented with those from the proctored quantitative versions.

Another measure taken in favor of legitimacy of quantitative item responses is offering inducements to the respondents (cf. page 26 for details). This is intended to encourage them to view the testing as an undertaking worth trying. For example, they will be sent a free score report to their inbox within 14 days of completion; there would be little point for them to seek external help from online reference materials only to obtain a report that does not reflect their proficiency level. Still, due to the technical infeasibility, their web browsers and task-switching cannot be controlled completely from the test server's end. On this account, the eighth limitation of this study must be acknowledged in that such a possibility of external help to their task engagement cannot be ruled out entirely.

Regarding the qualitative test versions ConCloze 2–4 and 7, the respondents are also offered inducements. The aim is to motivate them to cooperate in the testing (cf., e.g., White 2009: 82; Petre & Rugg 2010: 61; Schmitt 2010: 150 for rationale behind using inducement in research). However, rather than reward-seeking motivation, testing could generally involve washback effects including stress and anxiety (as in *test anxiety*) (cf., e.g., Messick 1996 for washback in language testing). Prior to high-stakes examinations, e.g., those for university admissions and TOEFL, test preparations such as revision and attending preparation courses can also be commonplace (Kim 2016). For research ethics and practicality, this study cannot replicate such washback effects that could be expected in high-stakes testing situations (cf. also pages 26f. for a restriction in detrimental effects caused to research participants). In light of this constraint, a final major limitation is that, in addition to the possibility of external help in the quantitative versions, the responses collected in the qualitative versions may or may not fully reflect the psychological states of task engagement that would be invoked otherwise. This means that while much effort has been made in this study to ensure population relevance of the examinee samples for minimal construct irrelevance, there could be no guarantee as to their test mentalities that would be represented in actual testing circumstances.

5.3 Recommendations for Future Research

5.3.1 On language testing

Formulaic language is routinely used expressions, entailing multi-word co-occurrences, e.g., fixed phrases, collocates (e.g., Sinclair 1991; Schmitt 2000; Sinclair 2003; Schmitt & Carter 2004; Schmitt *et al.* 2004b; Wood 2016). Accordingly, its concept may be deemed related to knowledge of word association and its subdomain of collocation, which is found to likely be tested in ConCloze (e.g., pages 141, 217). On this account, a first recommendation to future projects would be to administer a test on formulaic language and investigate how the scores from the test and ConCloze would actually be related. For example, Schmitt *et al.* (2004a) developed a contextualized test battery for formulaic sequences, one receptive item of which is exemplified in Figure 92 below (answer: *a. there's a good chance that*) (cf. also Wray 2002 for many varied but closed terms for formulaic language). Criterion-related validity evidence could be gathered for the ConCloze item type, such that whether or not knowledge of formulaic language would have any role to play in ConCloze engagement can be determined.

International debt

- Speaker A:** I've been watching the news report and they say that (11) _____ the international debts of poorer countries might be cancelled.
11. a. there's a good chance that
b. it seems to be happening that
c. the evidence is increasing that
d. people are thinking that
e. I DON'T KNOW

Figure 92 Schmitt et al.'s (2004a: 59) receptive format for a formulaic sequence

Currently there is a growing need for contextualized test formats that can systematically tap into multi-word lexical units (Read 2012). Yet, “there has been little systematic work [in the field] on how such lexical units should be assessed” (ibid.: 318). The few formats existing include Schmitt *et al.*'s (2004a) exemplified above. Accordingly, administering Schmitt *et al.*'s and ConCloze together would not only expand its criterion-related nomothetic net, but also potentially address the aforementioned scarcity by providing a new item type like ConCloze for the purpose. For example, if item responses to Schmitt *et al.*'s and ConCloze are found to share common

underlying factors (as in those looked for in the PCA, page 222), then an inference could be that the proficiency in vocabulary use tested in ConCloze is also likely to encompass knowing formulaic sequences. As such, ConCloze could be added to language testers' toolkit both for testing proficiency in vocabulary use and investigating score interpretations of other test formats that are hypothesized to tap into knowledge of formulaic language.

It is worth considering that the item format in Figure 92 above provides one context: *international debt*. In contrast, each ConCloze item presents multiple concordance lines as the context (as in Figure 1, page 1; cf. also page 4 for a distinction between *co-text* and *context*). This suggests that a single co-text is processed per item in Schmitt *et al.* (2004a) whereas multiple co-texts operate in ConCloze. In light of the scarcity in contextualized test formats, another inference could be that very few, if any, of the existing formats may draw sufficiently on differing aspects of formulaic language and word association. Accordingly, combining several related tests could produce a test suite that can deal with the multi-facets of formulaic language – another recommendation for future research with ConCloze taken into account. This thus means that considerable research opportunities could lie in both investigating the external aspect of ConCloze construct validity with each of those tests and analyzing their correlation patterns when administered alongside. For example, Gyllstad's (2007; 2009) COLLEX and COLLMATCH have been argued as relatively inappropriate as criterion tests for this study (page 53). Future studies could also challenge this view or seek to use ConCloze to supply criterion-related evidence for those tests.

Apart from the idea of forming a test suite for formulaic language, a third line of inquiry recommended is to proceed with investigating the test-score interpretation and use. In this study, a score meaning has been defined, indicating that a fundamental building block on the ConCloze item type has been provided. For example, the score meaning can be used as a basis for examining the value implications of ConCloze such as the stakeholders' perceptions when the test is administered locally for the first time (cf. also Figure 2, page 2 for the progressive validity model). Another example is when the score meaning is broken down into different aspects of proficiency in vocabulary use, which new ConCloze tests could be generated focusing on. A ConCloze battery would then be composed of the test from this study (representing the superordinate, generic aspect) and the newly generated ConCloze tests (representing the different aspects of the

construct). Administered together, they could form a test suite for proficiency in vocabulary use, assessing its multiple facets extensively. These two examples indicate that future studies can use the score meaning in this study as a foundation stone and the test as a measure.

This study investigates five out of six aspects of Messick's (1995) validity model (cf. page 16 for those aspects). The only aspect that is left unexplored is the consequential one. The first reason for omission is that it would be untenable to design a research project to look into the consequences of a test whose score meaning had been unknown. Also, because social consequences could be far-reaching, investigating them would deserve a separate study rather than being subsumed as part of exploring score meaning in this research. Related to continuing with ConCloze-score interpretation and use, a fourth line of inquiry is thus to investigate effects of the test on the societal level. For example, consequences of ConCloze implementation on local testing programs—e.g., the how-tos of item generation and banking in spite of limited computer literacy—may be focused on. Given the innovativeness of the item type, exploring such local consequences can be useful because it could be an onset of widespread corpus use in constructing language tests from the grassroot level, a rare social phenomenon that would warrant systematic investigations.

The network of word associations in the mental lexicon is complex (Schmitt & Meara 1997; Meara 2009). Given this, a fifth line of inquiry for future studies is to also look into how a ConCloze design impacts the measurement of different aspects of the construct proficiency. For example, one ConCloze test may be deliberately produced with sparse collocates in the concordance prompt, while the other has abundant collocates. If the two tests have the same set of options, then item difficulty could be investigated in relation to the effects of collocational presence/absence. If a strong presence of collocates leads to significant item easiness, it may be argued that words are stored and retrieved with their collocational relations prioritized. Investigating particular aspects of word association in ConCloze testing can be useful because it helps to build a case that its task content can be harnessed for measuring specialized domains of word knowledge (cf. also Kongsuwanakul 2014a for ideas of ConCloze applications).

Many English corpora divide text samples proportionately to different genres (e.g., Davies' (2008–) COCA). Given this, a final line of inquiry recommended for future

research could be to generate a ConCloze of specific genres such as newspaper and magazine. Administering a specific-genre ConCloze can be useful for assessing proficiency in vocabulary use for those studying journalism, for instance. Further, use of multi-word lexical units is a natural phenomenon not restricted to the English language (Schmitt & Carter 2004; Spöttl & McCarthy 2004; Schmitt 2010; Schmitt 2013); there exist corpora of languages other than English, e.g., Corpus del Español (Spanish). Developing a ConCloze of other languages and investigating the validity of its test-score interpretations would be useful for two reasons. First, the operationality of the tests would confer a universal utility for this item type. Secondly, finding a similar test interpretation would confer generalizability validity that cuts across languages in test-task content, another facet of construct measurement which has not been varied in this study.

5.3.2 On language acquisition and pedagogical practices

Tests almost always have impact, usually on teaching and learning. Accordingly, in addition to language testing (Section 5.3.1), future research on language learning and acquisition is also recommended. To begin with, the ConCloze proficiency is theorized to be acquired over the extent and intensity of exposure to English (page 272). Given this, a first line of inquiry could be experimental. For example, in schools where self-study centers are available, the students' behaviors as to the center use may be investigated when a ConCloze test is included as part of their final examinations. Studies may examine if the center use increases in frequency upon ConCloze incorporation, and whether the increase can lead to a better performance in the test and in general English proficiency. Investigating the impact of ConCloze on use of educational resources and on proficiency development can be useful because it would confirm that the proficiency level can be increased at the learners' own pace through exposure to English. Also, with the insights, policy makers would be able to allocate the resources to where they are most needed and benefit the students. On this account, investigating the washback effects of introducing a ConCloze format in an educational context could be recommended for future research.

It is worth reiterating that knowledge of lexical-semantics and knowledge of word association—two primary domains of knowledge found to be tested in ConCloze (page 264)—could potentially be deemed related to knowledge of formulaic language (page 283). This is in the sense that words and language components usually co-occurring with

them, e.g., collocation, colligation, semantic prosody, could become so commonly expected that they are regarded as formulaic. “[A] hallmark of the highest stages of language mastery” (Schmitt 2010: 145), formulaic language takes time to learn (Schmitt & Carter 2004: 13). Likewise, the ConCloze proficiency is also found to evolve over time (page 198). Accordingly, in addition to investigating the washback effects on use of educational resources, the item type could also expand its role in experimenting on how to efficiently acquire knowledge of formulaic language.

An example of an experimental study into efficiency of ConCloze in formulaic-language acquisition could be as follows. A ConCloze test may be generated and used as a pretest and posttest in investigating construct-relevant factors for formulaic-language acquisition. One control group of students may have only traditional methods of classroom practices, without any requirements of extensive self-study. One experimental group may be required to access their self-study center three times a week, two hours long each, without any exercises required. The other experimental group may be required to use the center similarly but only twice a week and with formulaic language-focused exercises that can be completed during the time they spend there. After one year, the three groups will then sit the ConCloze posttest and their gain scores will then be compared.

Considering the example of an experimental study, the aim of using ConCloze could be to examine the effectiveness of traditional teaching methods with that of incidental learning without active engagement and of deliberate learning with active engagement. The temporal difference between the two experimental groups could help to determine a covariate in acquisition: more incidental encounters and repetitions would be able to offset a lack of deliberate and actively-engaged learning of formulaic language (cf. also a review of the theory of lexical priming, pages 59ff.). Using ConCloze for measuring such gain scores in formulaic language can be useful because it could demonstrate which of the approaches would count as another “proactive, principled approach” generally needed for vocabulary learning (Schmitt 2010: 8). Insights into efficient teaching–learning approach(es) would enable the teachers and curriculum designers to plan the multi-word lexical units, the frequency of encounters needed, and the reinforcement of their primings for a best possible acquisition. Yet, these practices need to be assessed systematically, for example, in terms of their impact on teaching practices (e.g., more concordance use in class) and the impact on students’ use of corpora and concordancing (e.g., as a strategy to cope with the new item format). Accordingly,

investigating the impact of ConCloze on pedagogical practices can be another recommendation for future studies.

In addition to studying the impacts upon ConCloze introduction to school and classroom settings, future research may also investigate an indirect impact on those who do not have any experience with corpora. For example, when ConCloze tests are on the international or standardized level, language teachers who are not aware of corpus use may feel compelled to learn and integrate corpora to their lessons where possible (cf. also a usefulness of ConCloze as a new item type to the testing industry, pages 1f.). This could be particularly true when the situation involves the tests that the students need to get prepared for in the future, e.g., university-entrance examinations, TOEFL. On this account, a third line of inquiry for future studies could be to look into the social dimensions and an accelerating use of corpora for students' language learning when ConCloze is first introduced nationally or internationally.

A final area deserving further research is curriculum design. Presently, lexicogrammatical knowledge is viewed as significant for efficient language acquisition. This gives rise to the lexical syllabus, where lexis is given priority to, but not at the expense of, grammar or function (Sinclair & Renouf 1988; Greaves & Warren 2010; Szudarski 2013). In the syllabus, exposing the learners to authentic material and letting them draw conclusions about common patterns and chunking is a way of acquiring the knowledge (Sinclair 2004b). Because ConCloze also taps into lexical-semantic knowledge, with a construct-peripheral knowledge of grammatical structure, whether ConCloze could make a relevant assessment tool for the syllabus is another recommendation for future research.

Appendices

Appendix 1. ConCloze 1 Items

Question 1

1 balanced development, emphasizing tourism along the Alaska Highway corridor, and _____ among federal, territorial, and native planning initiatives. Although the results
2 to international organizations have initiated the practice of consultation and _____ among themselves, often without waiting for specific instructions from their
3 in Table 1 enabled each group to identify areas for _____ and joint ventures. The development of initial funding mechanisms through
4 in all of the other military-to-military contacts cited here, policy _____ has not always gone smoothly, and each side has attempted
5 must be related not only to case management, advocacy, and _____ of agency programs, but also to resource development in particular
6 for patients over 75 years old should be conducted, improved _____ with care assessment agencies is needed, and community services need
7 the teacher has the most concrete, practical ability to create _____ s, combinations, and mergers that are the basis for any idea

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|----------------|---------------|----------------|--------|
| A coordination | B integration | C organization | D work |
|----------------|---------------|----------------|--------|

Question 2

1 learn these or any teaching competency or skill is to _____ coursework with field experiences (Darling-Hammond, Hammerness, Grossman, Rust,
2 & Shulman, _____ed in complex ways. Initial forays into this question have yielded
3 composed of a large number of oscillators that are already _____ed to microenvironmental variations that may change significantly over small distances
4 contain few species, mostly microbiota, whose activity patterns are tightly _____ed together with a declaration of many years of productive married
5 of Rodolfo's unrepentant attitude and Leocadia's love mixed with fear, _____ed with a depressed postwar Laguna reservation economy subsequently prompted
6 the Laguna experience illustrates how the high level of unemployment _____ed with a depressed
7 she accrued them. In other cases, work schedules were tightly _____ed with economic conditions. One woman currently working in cosmetology explained
for the remainder of the decade. A 3.0% annual increase _____ed with projected population increases will raise military retirement costs to

All the lines above miss the same word. Which of the following should be that word?

A coupl(e)

B join

C pair

D suit

Question 4

1 theological detail ultimately undermines the properly interdisciplinary, nature of the _____ and results in a dominance of the social sciences in
2 their world, but that education in such a primary human _____ as music should be universal for our children in their
3 logical to go with your strengths and to structure the _____ in such a way as to guarantee success by anticipating
4 and coordination across divisions of the college must occur. This _____ is a campuswide effort that begins with a coordinating committee
5 or adventure travel. It is a tedious and time-consuming _____ that requires considerable inconsequential travel over vast expanses of territory.
6 devoted to social activity would leave less time for academic _____ s and therefore would likely have a negative impact on academic
7 being answered, but, as is always the case in scientific _____ s, more questions have arisen. We haven't yet found evidence for

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|------------------|--------------------------|-------------------|---------------------------|
| A attempt | B difficulty(/ie) | C endeavor | D fruitlessness(e) |
|------------------|--------------------------|-------------------|---------------------------|

Question 5

- 1 ideas. When one is improvising, it is not possible to _____ a passage that does not sound quite right, but in
 2 certain universal capacities that allow us to develop, sustain, and _____ a wide variety of cultural practices. For Diderot, Kant, and
 3 EPA. The act requires EPA to set and periodically to _____ national ambient air quality standards (NAAQS). These standards fix
 maximum
 4 President Bush and our governors for the year 2000, and _____ ed and continued by President Clinton, have put increased pressure on
 5 The Rules of Procedure implement the Standards. The Rules were _____ ed in 1989, and are amended from time to time. The
 6 imagination appears as prescriptive for theological method. He thereby substantially _____ s his earlier view that “deduction only is the instrument of
 7 or debunking the orthodox version of the past. Secondly, it _____ s, indeed transforms, the conventions and norms of historical fiction itself

All the lines above miss the same word. Which of the following should be that word?

- A** arrang(e) **B** chang(e) **C** consider **D** revis(e)

Question 6

- 1 including psychological distress and deteriorating spousal relations.
 2 An understanding of _____ adaptation in normative transitional circumstances (the birth of a child)
 3 et al., 1990). It is recommended that future research include the _____ and parental status of the coach when determining burnout levels.
 4 it is often the experience of healing and redemptive love. _____ Commitment # A final issue under this heading of the marriage
 5 and 3 are taken as "proxies for family dysfunction and _____ conflict." The results at the end of the time period
 6 than friendship or other important human relationships for in the _____ relationship it is understood that the entire life of both
 7 in potentially fruitful directions. Combat stress is, indeed, detrimental to _____ success, which implies that understanding the nature of military service
 ayete does not seem to be an autonomous form of _____ union. At the same time, however, the concession of women

All the lines above miss the same word. Which of the following should be that word?

- A discordant B marital C rightful D sexual

Question 7

- 1 bread and the wine offer by way of glimpsing human _____ and brokenness; for here precisely is where the Word and
- 2 Gary was groomed by men who picked up on his _____ and went on to sexually abuse and exploit him online.
- 3 to the present the nation became increasingly worried about economic _____ as a threat to national security. The time had come
- 4 located in a relatively active seismic region. The City's seismic _____ assessment of the local water system examines factors such as
- 5 and depression. Thus, empirical models studying the role of trait _____ in depression, and particularly in postpartum depression, should consider the
- 6 to address through their public work. It is also a[n] _____ that holds serious implications for gender roles and women's rights
- 7 have changed substantially since the mid-1950s, although the sense of _____ to Soviet nuclear attack may have increased; Soviet nuclear capabilities

All the lines above miss the same word. Which of the following should be that word?

A awareness **B** revelation **C** strength **D** vulnerability

Question 8

1 was not productive. The psychologist seemed to be trying to _____ a discussion between us regarding how and why Colleen found
 2 the ball instead of the player; the ball does not _____ a relationship -- it is not the thing on either side
 3 fields in the ways that economic backgrounds and ethnic/cultural heritage _____ children's development and schooling experiences (Frisby, 1999; Stoiber & Good,
 4 1998).
 5 _____ about musical behavior that tend to cloud our views of _____ ing musical learning, an expedition into areas regarded as alien territory
 6 explanations will be more satisfactory if they also identify processes _____ ing the effects of personal background on tolerance (Kohn, 1989). Future
 7 to cultural thought makes it a powerful mimetic referent that " _____ s all reflection and action upon the world" (Lock 1993). This
 8 daily activities and socialization. Language serves as a tool that _____ s the relationship between cognition and the environment. According to Vygotsky's

All the lines above miss the same word. Which of the following should be that word?

A condition B influenc(e) C mediat(e) D offer

Question 9

1 Nora has rejected God. By rejecting Torvald, by denying the _____ and dead father whose name she invoked with the forged
2 officials, the past century -- in which China has been largely _____ as an economic and military force -- was merely a historical
3 emotional adjustment were less strong for younger girls and were _____ for boys. # Parents not only influence their children, but children
4 calcium, depending on the species studied. The calcium layer is _____ from eggs produced by viviparous females, and the fiber layer
5 stature. By the mid-nineteenth century, anti-imperialist political thinking _____ from Western European intellectual debates, surfacing only rarely by way
6 In the group with previous surgery (N=27), the dizziness was _____ in 14 and was better in seven. The improvement in
7 acceptable in a democracy. As long as civilian control is _____, these states will be unlikely ever fully to consolidate democratic

All the lines above miss the same word. Which of the following should be that word?

A absent	B conspicuous	C minded	D present
----------	---------------	----------	-----------

Question 10

1 in a Comparative Context # This section examines the patterns of and mortality observed in the San Francisco Bay missions in a
 2 theory, is also consistent with Becker's (1976) microeconomic theory of and other economic approaches derived therefrom, which suggest that, all
 3 sexually active women in rural parts of China about their and risk of STDs and RTIs? What reproductive health problems
 4 in Iowa, the expectation based on the schedule of natural does not eliminate the effect of age. Indeed, a nearly
 5 # Some chemotherapy agents have long-term cardiac implications and there are issues linked with breast cancer treatments. There may also be
 6 countries. # "Between the end of World War II and 1970 rates rose virtually everywhere in the third world" (Abernathy 1993,
 7 law and order, and the natural environment suggesting life and . There is also the ocean, home of the Box Woman

All the lines above miss the same word. Which of the following should be that word?

A abundance **B** extension **C** fertility **D** increase

Question 11

1 precisely the sort of formal frames of reference that serve to ____ a broad reconsideration of academic freedom. He has made it
 2 other professionals such as physicians when all their efforts to ____ cooperation fail. Helping students differentiate where they can and should
 3 stimuli were used. Some involved social situations. For example, to ____ joy, parents were asked to play "peek-a-boo" with their infant.
 4 different racial and ethnic groups in our school. It also ____ed a genuine sense of understanding about culture and respect for
 5 origin. # In our patient 1, no neurologic deficit was ____ed pre- or postoperatively as a consequence of the tumor or
 6 narrators narrate and reflectors (usually) do not, Chatman's argument merely ____s a number of exceptions. # However, the fundamental misconception of (12)
 7 own sake but for the envy, admiration, or applause it ____s from others. But Washington differed from most in craving the

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|-----------------|-----------------|-------------------|---------------------|
| A elicit | B inform | C question | D respons(e) |
|-----------------|-----------------|-------------------|---------------------|

Question 12

1 traits, which are characterized by relative stability over time and across situations, positive state-like capacities are relatively more malleable and
 2 urbanization display three research perspectives: abstract frameworks that are universally but at the expense of substantive specificity; context-dependent frameworks that
 3 require the use of resources that require you clear the copyright laws before viewing and/or listening to them with your
 4 aim for a break, just for socialist reform, a reform in the new areas opened by the revolution and its
 5 still to be resolved. (n374) "Of 188 HAPs, 134 have measurement methods. Forty-three have likely methods. Eleven have potential or
 6 efficient and biologically appropriate. While this solution may not be to all common property resources, for those cases which may
 7 Court held that so long as a law is generally , and not discriminatory in being specifically aimed at religion, the

All the lines above miss the same word. Which of the following should be that word?

- A** applicable **B** included **C** suitable **D** true

Question 13

1 and Research. These feminists were primarily concerned with the communitarian _____ and the traditionalist question of identity, Who am I? In
 2 global touring, the accuracy or inaccuracy of one or another _____ concerning language classifications is less important than its contextual effect.
 3 less favorably than secular social civic, and recreational activities constitutes _____ discrimination, which is prohibited by the Free Speech Clause of
 4 students. Even if students do not actually disagree with a[n] _____ expressed on the discussion board, assuming the devil's advocate role
 5 to stop using the scientific method. We would think this _____ laughable because the risk of error is much greater when
 6 dealing with Colombia focused on program management and supported the _____ that both legislation and policy permit Colombia to use US
 7 positive attitude in the students towards the target language. The _____ that is to be given to the students on the

All the lines above miss the same word. Which of the following should be that word?

A appreciation B destruction C focus D viewpoint

Question 14

1 the violent break-up of Yugoslavia. It is not unreasonable to _____ a heightened sense of territoriality in these countries, translating into
 2 beliefs, hopes, fears, prejudices, and knowledge will lead us to _____ different authorial and narrative audiences as we infer these positions
 3 mid adolescence (test for interaction: $p < 0.001$). # Alternatively, one could also _____ that adolescents with high consumption of substances and behavioral problems
 4 of A.T. After a careful reading of the literature, they _____ ed and found three distinct clusters of behaviors identified as A.T:
 5 (Beck & Beck-Gernsheim, 2002; Giddens, 1992). In other words, we _____ ed teenage sexual behavior would be less influenced by membership in
 6 found that married people were happier than single people and _____ ed that marriage may draw persons who are already more satisfied
 7 role modeling of maternal smoking behavior. Notably, parental smoking is _____ ed to demonstrate pro-smoking norms and solidify pro-smoking attitudes. 17 18

All the lines above miss the same word. Which of the following should be that word?

- A formulat(e) B hypothesiz(e) C suggest D verify(/i)

Question 15

1 reign by virtue of public anachronism, insisting that they are _____ and relevant and politically indispensable, all the while acknowledging that
 2 by two cultures. This duality is not unrealistic but is _____ as African-Americans exist in and move between two cultures. Martin
 3 surmised that identity diffusion and moratorium may be viewed as _____ because both statuses provide adolescents opportunities to reflect on all
 4 each category and among categories. Diversity and heterogeneity capture the _____ capacity of a system, its ability to alter its composition
 5 just as well be a functionless byproduct of some other _____ feature of the human nervous system. Suppose, then, that for
 6 rather than competitiveness in the limit. # Proof The claim for _____ off-line adversaries follows from the fact that the two definitions
 7 approaches to develop a theory of optimal control of complex _____ systems. For management, this means in part asking how the

All the lines above miss the same word. Which of the following should be that word?

A adaptive **B** evolutionary **C** individual **D** responsive

Question 16

1 the freedom of the university and “the fight of any _____, subject to the normal rights of the heads of departments,
2 claims that courts may hesitate to reach. Though practitioners and _____s agree that the venerable statute remains effective after more than
3 on the Parisian landscape in June. But to many French _____s and museum professionals, the glitz and glamour of the opening
4 the steadily rising Muslim population caused immense concern among Hindu _____s and politicians who feared that the Hindus were destined to
5 this marginalized status, a peer subculture de-emphasized the importance of _____s by blaming external forces for their predicament, in summary, Adler
6 unrest, in other words, and not the good thinking of _____s that led to the creation of Native American studies (although,
7 becomes the principal contradiction of international politics. According to some _____s (Gilpin 1981, Organski 1958), the Chinese argue that international peace

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|------------|-------------|--------------|-----------|
| A academic | B professor | C researcher | D scholar |
|------------|-------------|--------------|-----------|

Question 17

1 in a body of diverse questions, it is important to _____ and cluster similar questions and connect them to the musical
2 Sinhalese # The Sinhala word jatiya (pl. jati) is used to _____ and distinguish all kinds of things, both animate and inanimate.
3 unreal extreme in behavior as a means to help us _____ and understand our own "trends" in philosophy and behavior. Its
4 age, education, and race. Participants self-reported their race/ethnicity and were _____ ed as African American; Asian or Pacific Islander; Hispanic or Latino;
5 (see Appendix A). Within each domain, research studies can be _____ ed as presenting strong evidence, promising evidence, or weak evidence that
6 for example, Health Benefits was one of the themes. Comments _____ ed under each theme were further examined to confirm, disconfirm, or
7 reproduced Figure 1. Their study itemizes the various techniques and _____ s them into techniques that verify the phases of processing and

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|----------------|-------------------|-------------------|--------|
| A categoriz(e) | B classify(/i/ie) | C identify(/i/ie) | D rank |
|----------------|-------------------|-------------------|--------|

Question 18

1 Liberalization of the economies in Northeast Asia is as yet
2 to be a problem. Unfortunately, the problem has been given
3 use of lower-frequency (e.g. annual) price information would result in
4 State Departments of Education. The Court found that there was
5 or cardiac death, however, indicates no benefit. Thus, ESPRIT provides
6 jobs, but employers report that despite large numbers of applicants,
7 on their own. Yet the food supply may still be

All the lines above miss the same word. Which of the following should be that word?

A available

B insufficient

Charge

D meager

Question 19

1 an increase of 1,020 deaths a year. # Scenario Three: The _____ continues only for the sale of spirits; wine and strong
 2 their credibility and effectiveness. If, for instance, there is a[n] _____ in the production of a good that is a vital
 3 the clerics, and for their supporters who possess a near _____ on the spoils of the revolution and the country's resources.
 4 and close friends. Food and fuel subsidies were removed, but _____ s and corruption continued. Examples of Suharto's inconsistencies included closing sixteen
 5 efficient companies, therefore, have usually been regarded not as evil _____ s but as national treasures. There is even a Japanese phrase,
 6 set of policies aimed to move swiftly to break up _____ s in specific sectors, such as coal, publishing and retail shops,
 7 private businesses to compete in sectors that were currently public _____ s. By February 1992, in the middle of the campaign for

All the lines above miss the same word. Which of the following should be that word?

- A abuse B capitalism C competition D monopoly(/ie)

Question 22

1 of Notre Daine. # Alison Flatau In addition to being an active mentor, _____ and educator, and receiving Director's Awards while at the NSF
 2 Disease Control and Prevention. # Carol Selman is a senior public health _____ at EHSB. She leads the Environmental Health Services Network (EHS-Net)
 3 in final form Goal setting 1st semester: Goals set after _____ meeting in December must be submitted by a date decided
 4 was incapable of responding logically to the crisis. Umar Juoro, a[n] _____ to Habibie and associate of Minister of Finance Ginandjar Kartasasmita, remarked
 5 10 chief executives from major research universities who served as _____s and champions for a 2005 American Council on Education study,"
 6 outlook. Making matters even more difficult. President Bush and his _____s have done little to dispel the notion that the administration
 7 spaceplace.nasa.gov. Alex Novati drew the illustrations.
 Thanks to Gene Schugart, Space Place _____, for activity concept and helpful advice. The article was provided

All the lines above miss the same word. Which of the following should be that word?

A advisor B chief C consultant D staff

Question 23

1 word value. Three definitions of intrinsic can be used to _____ an investigative framework for approximating intrinsic values and can be
2 verbal and nonverbal communication skills and to learn how to _____ his or her spare time and to take an active
3 with did not prove controversial. More problematic was how to _____ the government, how to divide power between the center and
4 zone of learning, the group process should to be carefully _____ ed and nurtured to ensure effective and efficient group work. Knowledge
5 to continue to improve the properties of materials by artificially _____ ing them on nanometer length scales and our ability to develop
6 epistolary confirmation of this idea. The picture itself quite literally _____ s a site for the I: in it Monet's place is
7 out of us a specific set of responses but also _____ s those responses. Working our way through a narrative results in

All the lines above miss the same word. Which of the following should be that word?

- | | | | | | | | |
|---|-----------|---|-------|---|-------------|---|-------------|
| A | arrang(e) | B | found | C | integrat(e) | D | structur(e) |
|---|-----------|---|-------|---|-------------|---|-------------|

Question 24

1 the marketplace. We are often asked to rejoice that the 4% success rate is so high. If low success is
 2 culture to culture, however, subject to concepts of individualism, the aesthetic, and a host of social or ritual beliefs particular
 3 way to a salon. There are many methods, some even almost since the birth of photography. The first method is
 4 instruction. We failed to appreciate how deeply entrenched were the attitudes about what constituted appropriate social studies instruction for the
 5 quarrel about their theoretical compatibility, these were two sets of ideas that were very powerful in American management. The first
 6 focusing entirely on the dining-room dimension of language. Thus the linguistic point of view ends up indistinguishable from that of
 7 you think you have made an important discovery counter to opinion? Over lunch a number of years ago, one senior

All the lines above miss the same word. Which of the following should be that word?

- A current B indicative C prevailing D standardized

Question 25

1 Asia is clearly more critical for China because of its _____ and strategic relevance. However China's economic diplomacy in the region
2 distinct polyploid genotypes of independent origin that are in close _____ can cross with one another, leading to even more genetic
3 Stadium if they bore increased costs because of (a) the _____ of their homes to the stadium, (b) homes located behind
4 credit from any of these people, no matter their actual _____ to his informal personal trade network, which also includes Korean,
5 deeper investigation into their moral characteristics and criminal impulses. Criminality, _____, and the legibility of race within colonial and postcolonial theory
6 requires disclosing the ways in which the mutual dependencies and _____s created by globalization make it increasingly impossible to designate some
7 within a single mesocosm in a way that retains realistic _____s, interactions, and relationships. This has led some researchers to design

All the lines above miss the same word. Which of the following should be that word?

- A distance
- B immediacy(/ie)
- C proximity(/ie)
- D region

Question 26

- 1 and the subsequent controversial rise and fall of Rajneeshpuram helped _____ a climate of public hostility (and ultimately judicial hostility) to
 2 weeds is therefore linked, in part, to early efforts to _____ land use on the ranchettes. The market explanation does not
 3 problems. # As for immediate security issues, Washington and Beijing should _____ their efforts to secure the stability of the Korean peninsula.
 4 in women's and men's labor and livelihoods as cotton production _____ ed are read against processes of cultural production, particularly the production
 5 1966, which brought the Baath party to power, Egyptian-Syrian rivalry _____ ed, and support for the Palestinian cause was, as usual, one
 6 to taboo preserves human resources in a period when work _____ s and labor demands are increasingly placed on people who already
 7 sin of terrorism may be that it apes or even _____ s the worst aspects of the oppressive regimes from which the

All the lines above miss the same word. Which of the following should be that word?

- A intensify(/i/ie) B persecut(e) C sens(e) D strengthen

Question 27

1 them to support HSO staff by ensuring that they are _____ about ethics issues, have learned a methodology for addressing the
 2 group to be primarily older and white and more highly _____ and less religiously affiliated than the general population (Kamakahi, Cossman,
 3 provided them with enough information and experience to make a[n] _____ decision about what they need to do to achieve specific
 4 understandings of appropriate female roles contributing to negative stereotypes of _____ girls. In this section, symbolic boundaries are further explored, particularly
 5 the traditional economic assets of a city (such as a[n] _____ labor force, low taxes, and good business climate) to the
 6 the tales of desperate battles and the whiff of gunpowder. _____ society picks over the descriptions of strange lands, their plants
 7 and their parents. The parents tended to be Caucasian, highly _____, upper income status, and from a wide variety of cities

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|---------------------|-------------------|-------------------|---------------------|
| A articulate | B cultured | C educated | D scientific |
|---------------------|-------------------|-------------------|---------------------|

Question 28

1 However, the droughts in 1969-73 and 1984 have considerably diminished and pastures, thereby weakening some parental, elderly, and chiefly leverage.
2 island laboratories, where they search for new weapons against this disease that threatens no consumer's health, but every consumer's pocketbook.
3 available in the state food markets, and anyone caught slaughtering illegally may be imprisoned for up to four years, but
4 for the spread of sleeping sickness. 4 The restocking of in Teso has been identified as a central activity for
5 the local lumber yard, they could contract to ship their to South St. Paul, and they could deal with local grain
6 injure his family, and then escape with some of his . When they return, Farooq has gathered friends to help him
7 acquired through purchased conservation easements while ranchers continue to raise . This approach has become increasingly popular, especially among NGOs such

All the lines above miss the same word. Which of the following should be that word?

A breeding

B creature

C livestock

D stray

Question 29

1 of states. It has become fashionable among political scientists to _____ a causal link between democracy and peace, extrapolating from the
2 how she uses her body, as Jardine would say, to _____ an image different from that of the traditional image of
3 This puts the authors in a paradox: how can they _____ students as clients, with the power of choice, and also
4 value individual expression less than low-context cultures, some researchers have _____ ed a relationship between culture and CA. Gudykunst and Ting-Toomey (1988),
5 perspective) with a specific cultural group. Hence, if the propositions _____ ed by the theory are observed, then it is assumed that
6 political violence and another led by the Partido Popular, which _____ ed that the laying down of arms by ETA would actually
7 from an articulation of the booster discourse. Specifically, the theory _____ s a set of common strategies that boosters follow to "grow"

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|------------|-----------|---------|-----------|
| A argue(e) | B explain | C posit | D suggest |
|------------|-----------|---------|-----------|

Question 30

- 1 community as a full-fledged member. # North Korea's soft landing, or ____ adoption of a market economy and liberal democracy, is desirable
 2 equilibrium -- and that change does not usually occur by imperceptibly ____ alteration of entire species but rather by isolation of small
 3 to make any major impacts for about 20 years. But ____ and visible improvements, both environmental and economic, can be realized
 4 males partially deviated from this orientation: One expressed a more ____ approach to dating and sex, and another voiced a more
 5 faster pace of environmental improvement and those arguing for more ____ change. # Motor vehicle use, already discouraged through high registration fees
 6 predisposition to Huntington's disease, in which "afflicted individuals
 experience a[n] ____ loss of control over their muscles, body, and mind." Because
 7 what should worry us most is our own inwardness, our ____ withdrawal from the rest of society. We participate less in

All the lines above miss the same word. Which of the following should be that word?

- A** ascending **B** gradual **C** improved **D** rapid

Question 31

in Uruguay and cleaned and released in 2002, ran into United Nations, under U.S. leadership, fought a modern war over and Shea. He also serves on the Underground Storage Tank become more variable due to the greenhouse effect. Without cheap there are 4,500 companies and establishments that supply the national Asia and the Transcaucasus in exchange for cotton, gas, and been successfully developed and evaluated for the determination of common

All the lines above miss the same word. Which of the following should be that word?

A chemical B fuel C petroleum D wax

Question 32

1 although far fewer than the stories they must organize, still a fairly ample range. For example, creation stories may be
 2 move toward an electronic, worldwide market. That market will probably all time zones and ignore all national boundaries. It will
 3 and children's literature. Brenner's major translations of creative works ostensibly the years 1930 to 1944. 14 Three are novels by Mexican
 4 Marshall went on to enjoy a major league career that ed 14 years. He set several pitching records in the 1970s
 5 deep commitment to embodied inquiry. Two teachers, whose lives have ed the century, give a particular flavor of the kind of
 6 Mexico; the Very Long Baseline Array, with ten eighty-two-foot dishes ing 5,000 miles from Hawai'i to the Virgin Islands; and the
 7 there is no bright-line test for determining nuisance. Rather, nuisance s a broad spectrum. # In discussing nuisance this Comment has established

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|--------------------|------------------|------------------|------------------|
| A includ(e) | B prolong | C span(n) | D stretch |
|--------------------|------------------|------------------|------------------|

Question 33

- 1 the presence of administrators in the role of evaluators. A[n] _____ alternative is the inclusion of teachers in the evaluation process.
 2 as good stories usually are, the chief's saga is internally _____ and logical; a headman whose nation's territory included a highly
 3 all such inferences, however, our result is only as _____ as the model used for imputing the missing data. We
 4 acknowledging their contribution to his projects -- consistent
 5 character motivation and _____ character psychology -- which he disguised or "smoked" within a commentative,
 6 a leader has been in power. Deductively, this hypothesis is _____ for a number of reasons. # First, those leaders who face
 7 unaffordable form of old-fashioned social democracy. This absence of a[n] _____ progressive counter-narrative is unhealthy, because competition is good for
 intellectual
 I propose for Gudlaugr's extraordinary appearance seems to me more _____ than the secular reading of this scene which Sayers suggests,

All the lines above miss the same word. Which of the following should be that word?

- A explanatory B ingenuous C plausible D satisfactory

Question 34

1 Syrian fortress of Shaizar, Usamah had lost his right to ____ and had become a cavalier -- a faris -- ready to serve
 2 the price of such politics would be wars of dynastic ____ and internecine violence waged in order to determine who would
 3 to the actual king to avoid impending wars over the ____ by choosing an heir immediately. Iberia, nevertheless, is also a
 4 will discard all procedure at all cost to change the ____ for the highest power. # Deng Xiaoping is an emperor without the
 5 however, whether anyone in Washington has thought of aiding Oman's ____ in the manner that Britain did decades ago to ensure
 6 not been held in high esteem by most Thais. His ____ to the throne could undermine the mystique of the monarchy,
 7 democracies. Democracies enhance political stability by establishing clear mechanisms for _____. This allows them to adapt smoothly to the death or

All the lines above miss the same word. Which of the following should be that word?

A appointment B empire C position D succession

Question 35

- 1 the school counselor of their suspicions, though this may directly _____ applicable state laws, which often explicitly list teachers, administrators, and
school
- 2 to ceremonies of exorcism that are injurious to animals and _____ Buddhist norms (Gombrich and Obeyesekere 1988:142). Third, some Buddhist
ascetics
- 3 successfully navigate the free market. The actions of prosperous entrepreneurs _____ the standard of egalitarianism previously maintained by the state. This
governmental dysfunction and rampant institutionalized corruption.
- 4 These sweeping statements are _____ ed by concrete examples of effective government action and achievements. Thus,
5 were virtually unknown, even in South Africa, where their existence _____ ed the apartheid myth that black peoples there had arrived only
6 give some of their wealth to those in absolute poverty _____ s their right to their private property. In talking about the
7 dramatizes the characters' difficult identification with an image that continually _____ s what they see as their "real" image. Being subjected to

All the lines above miss the same word. Which of the following should be that word?

- A** confound **B** contradict **C** propos(e) **D** refut(e)

Question 36

when abused as a child and therefore it is his
and the Holocaust as a comparative framework and not other
explore not the facts of her rape but the enduring
to the point where it has been fully experienced, the
to change. These do not automatically disappear because the original
reveal memory, insisting that the telling and the visualizing of
those evenings which imprisonment assassinated (27).
Admittedly, imprisonment can be
, and many prisoners feel distraught, disoriented, and experience a sense

All the lines above miss the same word. Which of the following should be that word?

Anxious

B conditioned

C disagreeable

D traumatic

Question 37

1 their fears -- the fear that computers will mean their own and join to help encourage an area that could strengthen
 2 the number of back disorders. # However, injury reduction and eventual can only be accomplished through the joint efforts of employees
 3 fewer providers, such as E.I. du Pont de Nemours and Co.'s highly publicized 1992 of 90 percent of its outside firms that left just
 4 definition and clarification of hazing to ensure proper enforcement and of all negative behaviors. # In the fall of 2007, a
 5 block grant in the 1980s when the Reagan administration emphasized of governmental agencies and regulations. Moreover, Flemming was the first
 6 authority common to all treaty committees. The Committee on the of Racial Discrimination has benefitted from the international system's strong
 7 sword and some other cool weapons especially designed for vampire , and the fight proceeds, more or less, as one might

All the lines above miss the same word. Which of the following should be that word?

A completion B elimination C process D removal

Question 38

1 but throughout eastern Kansas and even western Missouri. Historians often _____ events after 1861, though technically a part of the region's
 2 _____ if the Gaussians are sufficiently well separated then one can usually _____ together all the samples originating from the same Gaussian. As
 3 as having no experience of same-sex behavior. # Respondents were then _____ ed according to their sexual experiences and sexual attractions based on
 4 all goods in an individual household, though goods were frequently _____ ed in categories or only given the barest description by valuers.
 5 noted the sources of their information and reactions. Sources were _____ ed in the following six categories: friends, parents, teachers, media, books
 6 analysis encompasses a number of different algorithms and methods for _____ ing objects of similar kind into respective categories to develop taxonomies.
 7 advantages to using this approach outweigh disadvantages, the scheme often _____ s together individual occupations into broader categories for coding purposes. For

All the lines above miss the same word. Which of the following should be that word?

A divid(e)

B gather

C group

D select

Question 39

1 call "first-person authority," wherein an individual is said to have _____ access to his own inner life, weakens to the point
 2 not usually thought of as stigmatized because they are seen _____ and unlike most other stigmas they chose to be athletes
 3 the whole people inevitably changes to actual ownership by the _____ class. Because the people have had no power for such
 4 owners -- of revolutionary nation-building. # While I have stressed the men's _____ position in the village and the revolution, I must underline
 5 policy standpoint, educational practices often have the effect of favoring _____ students and hindering the educational opportunities of students of color.
 6 comes alive with each moment of teaching that I am _____ to receive. This is a central role for me as
 7 expression that is acknowledged as religiously, politically, traditionally, or morally "_____" and the compelling word that is "half-ours and half-someone else's,"

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|-----------------------|------------------|----------------------|---------------------|
| A aristocratic | B divided | C prerogative | D privileged |
|-----------------------|------------------|----------------------|---------------------|

Appendix 2. ConCloze 2–4 Items

Question 1

1 theological detail ultimately undermines the properly interdisciplinary, nature of the _____ and results in a dominance of the social sciences in
2 their world, but that education in such a primary human _____ as music should be universal for our children in their
3 logical to go with your strengths and to structure the _____ in such a way as to guarantee success by anticipating
4 and coordination across divisions of the college must occur. This _____ is a campuswide effort that begins with a coordinating committee
5 or adventure travel. It is a tedious and time-consuming _____ that requires considerable inconsequential travel over vast expanses of territory.
6 devoted to social activity would leave less time for academic _____ s and therefore would likely have a negative impact on academic
7 being answered, but, as is always the case in scientific _____ s, more questions have arisen. We haven't yet found evidence for

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|-----------|-------------------|------------|--------------------|
| A attempt | B difficulty(/ie) | C endeavor | D fruitlessness(e) |
|-----------|-------------------|------------|--------------------|

Question 2

1 traits, which are characterized by relative stability over time and across situations, positive state-like capacities are relatively more malleable and
 2 urbanization display three research perspectives: abstract frameworks that
 are universally but at the expense of substantive specificity; context-dependent frameworks that
 3 require the use of resources that require you clear the copyright laws before viewing and/or listening to them with your
 4 aim for a break, just for socialist reform, a reform in the new areas opened by the revolution and its
 5 still to be resolved. (n374) "Of 188 HAPs, 134 have measurement methods. Forty-three have likely methods. Eleven have potential or
 6 efficient and biologically appropriate. While this solution may not be to all common property resources, for those cases which may
 7 Court held that so long as a law is generally , and not discriminatory in being specifically aimed at religion, the

All the lines above miss the same word. Which of the following should be that word?

A applicable

B suitable

C true

Question 3

1 It is not unreasonable to ____ a heightened sense of territoriality
2 knowledge will lead us to ____ different authorial and narrative audiences
3 p < 0.001). # Alternatively, one could also ____ that adolescents with high consumption
4 reading of the literature, they ____ ed and found three distinct clusters
5 1992). In other words, we ____ ed teenage sexual behavior would be
6 happier than single people and ____ ed that marriage may draw persons
7 behavior. Notably, parental smoking is ____ ed to demonstrate pro-smoking norms and

All the lines above miss the same word. Which of the following should be that word?

- A formulat(e) B hypothesiz(e) C suggest D verify(/i)

Question 4

1 commitments in their lives such as family, employment, and other activities and hobbies that capture their interest. The college players
2 frequent, use. If their use of these drugs is often and associated with leisure activities, then it is likely that
3 appear to yield different results from comparisons between competitive and athletes. # Furthermore, an examination of specific categories of superstitious
4 most marinas. # A large number of respondents also commented that boaters are regarded as "fat cats" and are being "picked
5 defined polydrug use as the use of 3 or more drugs in the past 90 days (meth-amphetamine, cocaine, crack, MDMA
6 on how well the library was meeting users' educational and needs so improvements could be made. They agreed to complete
7 to the United States, where most bicycles sold are for use, most of the 105 million bicycles sold each year worldwide

All the lines above miss the same word. Which of the following should be that word?

- A avian B fluffy C mechanical D recreational

Question 5

- 1 However, the droughts in 1969-73 and 1984 have considerably diminished _____ and pastures, thereby weakening some parental, elderly, and chiefly leverage.
 2 island laboratories, where they search for new weapons against this _____ disease that threatens no consumer's health, but every consumer's pocketbook.
 3 trade allowed Chiricahua and other Apache to exchange their stolen _____ from raids in Sonora and Chihuahua for the food, guns,
 4 available in the state food markets, and anyone caught slaughtering _____ illegally may be imprisoned for up to four years, but
 5 for the spread of sleeping sickness. 4 The restocking of _____ in Teso has been identified as a central activity for
 6 (Landau et al. 1995), which can lead to large increases in _____ numbers and, as a result, to the occurrence of localized
 7 walk. Even though the animal is likely to recover, the _____ producer's worries are hardly over, notes microbiologist Cecelia A. Whetstone, head
 8 the local lumber yard, they could contract to ship their _____ to South St. Paul, and they could deal with local grain
 9 acquired through purchased conservation easements while ranchers continue to raise _____. This approach has become increasingly popular, especially among NGOs
 10 injure his family, and then escape with some of his _____. When they return, Farooq has gathered friends to help him
 such

All the lines above miss the same word. Which of the following should be that word?

- A** breeding **B** creature **C** livestock **D** stray

Appendix 3. ConCloze 5 Items

Question 1

1 ideas. When one is improvising, it is not possible to _____ a passage that does not sound quite right, but in
2 certain universal capacities that allow us to develop, sustain, and _____ a wide variety of cultural practices. For Diderot, Kant, and
3 EPA. The act requires EPA to set and periodically to _____ national ambient air quality standards (NAAQS). These standards fix
4 _____ maximum
5 President Bush and our governors for the year 2000, and _____ ed and continued by President Clinton, have put increased pressure on
6 The Rules of Procedure implement the Standards. The Rules were _____ ed in 1989, and are amended from time to time. The
7 imagination appears as prescriptive for theological method. He thereby substantially _____ s his earlier view that “deduction only is the instrument of
or debunking the orthodox version of the past. Secondly, it _____ s, indeed transforms, the conventions and norms of historical fiction itself

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|---------|----------|------------|-----------|
| A grill | B hyp(e) | C revis(e) | D warrant |
|---------|----------|------------|-----------|

Question 2

- 1 bread and the wine offer by way of glimpsing human _____ and brokenness; for here precisely is where the Word and
 2 Gary was groomed by men who picked up on his _____ and went on to sexually abuse and exploit him online.
 3 to the present the nation became increasingly worried about economic _____ as a threat to national security. The time had come
 4 located in a relatively active seismic region. The City's seismic _____ assessment of the local water system examines factors such as
 5 and depression. Thus, empirical models studying the role of trait _____ in depression, and particularly in postpartum depression, should consider the
 6 to address through their public work. It is also a[n] _____ that holds serious implications for gender roles and women's rights
 7 have changed substantially since the mid-1950s, although the sense of _____ to Soviet nuclear attack may have increased; Soviet nuclear capabilities

All the lines above miss the same word. Which of the following should be that word?

- A advancement B environment C integrity D vulnerability

Question 3

precisely the sort of formal frames of reference that serve to
other professionals such as physicians when all their efforts to
stimuli were used. Some involved social situations. For example, to
different racial and ethnic groups in our school. It also
century later (1944). Truly enough, Myrdal's analysis was sponsored and
computers, is a fine example of an ear-training program that
own sake but for the envy, admiration, or applause it

All the lines above miss the same word. Which of the following should be that word?

A design **B** elicit **C** govern **D** seek(/sought)

Question 4

1 noted some qualitative similarities across countries: we could not reject _____ behavior; investment tends to be the most significant variable for
 2 their credibility and effectiveness. If, for instance, there is a[n] _____ in the production of a good that is a vital
 3 the clerics, and for their supporters who possess a near _____ on the spoils of the revolution and the country's resources.
 4 and close friends. Food and fuel subsidies were removed, but _____ s and corruption continued. Examples of Suharto's inconsistencies included closing sixteen
 5 efficient companies, therefore, have usually been regarded not as evil _____ s but as national treasures. There is even a Japanese phrase,
 6 set of policies aimed to move swiftly to break up _____ s in specific sectors, such as coal, publishing and retail shops,
 7 private businesses to compete in sectors that were currently public _____ s. By February 1992, in the middle of the campaign for

All the lines above miss the same word. Which of the following should be that word?

- A** accumulation **B** information **C** monopoly(/ie) **D** research(e)

Question 5

- 1 the school counselor of their suspicions, though this may directly _____ applicable state laws, which often explicitly list teachers, administrators, and
school
- 2 _____ to ceremonies of exorcism that are injurious to animals and _____ Buddhist norms (Gombrich and Obeyesekere 1988:142). Third, some Buddhist
ascetics
- 3 successfully navigate the free market. The actions of prosperous entrepreneurs _____ the standard of egalitarianism previously maintained by the state. This
governmental dysfunction and rampant institutionalized corruption.
- 4 These sweeping statements are _____ ed by concrete examples of effective government action and achievements. Thus,
5 _____ were virtually unknown, even in South Africa, where their existence _____ ed the apartheid myth that black peoples there had arrived only
6 _____ give some of their wealth to those in absolute poverty _____ s their right to their private property. In talking about the
7 dramatizes the characters' difficult identification with an image that continually _____ s what they see as their "real" image. Being subjected to

All the lines above miss the same word. Which of the following should be that word?

- A contradict B provid(e) C requir(e) D suggest

Question 6

1 resulted in a purchase or _____ action at the advertiser's site.
 2 well suited to provide the _____ baseline of AIS-selection information. As
 3 a list of choices. The _____ category may then be selected
 4 The opportunity to claim a[n] _____ identity was operationalized as winning
 5 coils that can generate the _____ magnetic flux. These methods also
 6 between the status quo and the _____ reference point of the community.
 7 with respect to current and _____ weight ($F = 4.85$, $p = .0134$). Tests of

All the lines above miss the same word. Which of the following should be that word?

A amorous **B** desired **C** gained **D** passionate

Question 7

1 effects of life expectancy on and family planning are statistically
 2 Becker's (1976) microeconomic theory of and other economic approaches derived
 3 parts of China about their and risk of STDs and
 4 on the schedule of natural does not eliminate the effect
 5 cardiac implications and there are issues linked with breast cancer
 6 These scholars deny that African levels come from local values,
 7 natural environment suggesting life and . There is also the ocean,

All the lines above miss the same word. Which of the following should be that word?

A abundance **B** extension **C** fertility **D** increase

Question 8

1 by ensuring that they are _____ about ethics issues, have learned
2 and white and more highly _____ and less religiously affiliated than
3 and experience to make a[n] _____ decision about what they need
4 contributing to negative stereotypes of _____ girls. In this section, symbolic
5 a city (such as a[n] _____ labor force, low taxes, and
6 and the whiff of gunpowder. _____ society picks over the descriptions
7 tended to be Caucasian, highly _____, upper income status, and from

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|---------------------|-------------------|-------------------|---------------------|
| A articulate | B cultured | C educated | D scientific |
|---------------------|-------------------|-------------------|---------------------|

Question 9

1 and therefore it is his childhood experience that is to
 2 comparative framework and not other contexts such as Rwanda or
 3 her rape but the enduring effects of it on her
 4 has been fully experienced, the event moves from the present
 5 around two points: first, that events have usually been those
 6 was clear that the woman's experience during World War II
 7 of coping with and expressing issues and experiences. The individual

All the lines above miss the same word. Which of the following should be that word?

A anxious **B** conditioned **C** disagreeable **D** traumatic

Question 10

1 individual is said to have _____ access to his own inner
 2 stigmatized because they are seen _____ and unlike most other stigmas
 3 to actual ownership by the _____ class. Because the people have
 4 seems to have been, this _____ group is also more casual
 5 have the effect of favoring _____ students and hindering the educational
 6 of teaching that I am _____ to receive. This is a
 7 religiously, politically, traditionally, or morally "_____" and the compelling word that

All the lines above miss the same word. Which of the following should be that word?

A aristocratic

B divided

C prerogative

D privileged

Question 11

1 balanced development, emphasizing tourism along the Alaska Highway corridor, and _____ among federal, territorial, and native planning initiatives. Although the results
 2 to international organizations have initiated the practice of consultation and _____ among themselves, often without waiting for specific instructions from their
 3 _____ in Table 1 enabled each group to identify areas for _____ and joint ventures. The development of initial funding mechanisms through
 4 _____ in all of the other military-to-military contacts cited here, policy _____ has not always gone smoothly, and each side has attempted
 5 _____ must be related not only to case management, advocacy, and _____ of agency programs, but also to resource development in particular
 6 _____ for patients over 75 years old should be conducted, improved _____ with care assessment agencies is needed, and community services need
 7 _____ the teacher has the most concrete, practical ability to create _____ s, combinations, and mergers that are the basis for any idea

All the lines above miss the same word. Which of the following should be that word?

- | | | |
|-----------------------|-----------------------|---------------|
| A coordination | B organization | C work |
|-----------------------|-----------------------|---------------|

Question 12

1 theological detail ultimately undermines the properly interdisciplinary, nature of the _____ and results in a dominance of the social sciences in
 2 their world, but that education in such a primary human _____ as music should be universal for our children in their
 3 logical to go with your strengths and to structure the _____ in such a way as to guarantee success by anticipating
 4 and coordination across divisions of the college must occur. This _____ is a campuswide effort that begins with a coordinating committee
 5 or adventure travel. It is a tedious and time-consuming _____ that requires considerable inconsequential travel over vast expanses of territory.
 6 devoted to social activity would leave less time for academic _____ s and therefore would likely have a negative impact on academic
 7 being answered, but, as is always the case in scientific _____ s, more questions have arisen. We haven't yet found evidence for

All the lines above miss the same word. Which of the following should be that word?

- A difficulty(/ie) B endeavor C fruitlessness(e)

Question 13

reign by virtue of public anachronism, insisting that they are
by two cultures. This duality is not unrealistic but is
surmised that identity diffusion and moratorium may be viewed as
each category and among categories. Diversity and heterogeneity capture the
just as well be a functionless byproduct of some other
rather than competitiveness in the limit. # Proof The claim for
approaches to develop a theory of optimal control of complex

All the lines above miss the same word. Which of the following should be that word?

A adaptive

B evolutionary

Individual

Question 14

- 1 the presence of administrators in the role of evaluators. A[n] _____ alternative is the inclusion of teachers in the evaluation process.
 2 as good stories usually are, the chief's saga is internally _____ and logical; a headman whose nation's territory included a highly
 3 all such inferences, however, our result is only as _____ as the model used for imputing the missing data. We
 4 acknowledging their contribution to his projects -- consistent _____ character psychology -- which he disguised or "smoked" within a commentative,
 5 to us. Thanks to brain science, we now have a[n] _____ explanation for such experiences. We didn't enjoy the book that
 6 unaffordable form of old-fashioned social democracy. This absence of a[n] _____ progressive counter-narrative is unhealthy, because competition is good for
 intellectual
 7 I propose for Gudlaugr's extraordinary appearance seems to me more _____ than the secular reading of this scene which Sayers suggests,

All the lines above miss the same word. Which of the following should be that word?

- A** ingenuous **B** plausible **C** satisfactory

Question 15

1 Syrian fortress of Shaizar, Usamah had lost his right to ____ and had become a cavalier -- a faris -- ready to serve
 2 the price of such politics would be wars of dynastic ____ and internecine violence waged in order to determine who would
 3 to the actual king to avoid impending wars over the ____ by choosing an heir immediately. Iberia, nevertheless, is also a
 4 will discard all procedure at all cost to change the ____ for the highest power. # Deng Xiaoping is an emperor without the
 5 however, whether anyone in Washington has thought of aiding Oman's ____ in the manner that Britain did decades ago to ensure
 6 not been held in high esteem by most Thais. His ____ to the throne could undermine the mystique of the monarchy,
 7 democracies. Democracies enhance political stability by establishing clear mechanisms for _____. This allows them to adapt smoothly to the death or

All the lines above miss the same word. Which of the following should be that word?

A empire B position C succession

Question 16

- 1 including psychological distress and deteriorating spousal relations.
 2 An understanding of _____ adaptation in normative transitional circumstances (the birth of a child)
 3 et al., 1990). It is recommended that future research include the _____ and parental status of the coach when determining burnout levels.
 4 both sides, "Are these mail-order brides jumping on the long-distance _____ band wagon just to get a passport to the promised
 5 inspirational readings offered women optimism and motivated affirmation _____ of the _____ bond as a "covenant" between God and both members of
 6 and 3 are taken as "proxies for family dysfunction and _____ conflict." The results at the end of the time period
 7 for older but not for younger adolescents. The associations between _____ distress of parents and emotional adjustment of adolescents did not
 8 since marriages are contracted by families and not the individuals. _____ instability is in the same league as poverty. It leads
 9 than friendship or other important human relationships for in the _____ relationship it is understood that the entire life of both
 10 in potentially fruitful directions. Combat stress is, indeed, detrimental to _____ success, which implies that understanding the nature of military service
 ayete does not seem to be an autonomous form of _____ union. At the same time, however, the concession of women

All the lines above miss the same word. Which of the following should be that word?

- A bridal B marital C rightful

Question 17

1 proceed. This gives an assigning lawyer the benefit of another and also a chance to catch any misperceptions on the
 2 and Research. These feminists were primarily concerned with the communitarian and the traditionalist question of identity. Who am I? In
 3 global touring, the accuracy or inaccuracy of one or another concerning language classifications is less important than its contextual effect.
 4 decision to prohibit political advertising on bus placards amount to discrimination when in the course of applying this policy it
 5 less favorably than secular social civic, and recreational activities constitutes discrimination, which is prohibited by the Free Speech Clause of
 6 students. Even if students do not actually disagree with a expressed on the discussion board, assuming the devil's advocate role
 7 to stop using the scientific method. We would think this laughable because the risk of error is much greater when
 8 in the wake of Ann Banfield, free indirect style encodes the of the character, that is, is a representation of his
 9 dealing with Colombia focused on program management and supported the that both legislation and policy permit Colombia to use US
 10 positive attitude in the students towards the target language. The that is to be given to the students on the

All the lines above miss the same word. Which of the following should be that word?

- | | | |
|----------------|----------------------|--------------------|
| A focus | B perspective | C viewpoint |
|----------------|----------------------|--------------------|

Question 18

Liberalization of the economies in Northeast Asia is as yet
to be a problem. Unfortunately, the problem has been given
use of lower-frequency (e.g. annual) price information would result in
State Departments of Education. The Court found that there was
or cardiac death, however, indicates no benefit. Thus, ESPRIT provides
number of qualified braille transcribers, to examine the issues of
overall health status (Mellor & Lindeman, 1998). Factors such as
jobs, but employers report that despite large numbers of applicants,
on their own. Yet the food supply may still be
to the home market. Similarly, if such diversion is still

All the lines above miss the same word. Which of the following should be that word?

A available

B insufficient

C meager

Question 19

1 the marketplace. We are often asked to rejoice that the ____ 4% success rate is so high. If low success is
 2 culture to culture, however, subject to concepts of individualism, the ____ aesthetic, and a host of social or ritual beliefs particular
 3 way to a salon. There are many methods, some even ____ almost since the birth of photography. The first method is
 4 instruction. We failed to appreciate how deeply entrenched were the ____ attitudes about what constituted appropriate social studies instruction for the
 5 need to be put off by e-discovery. In fact, the ____ court rulings have created standards close to the Federal Rules
 6 quarrel about their theoretical compatibility, these were two sets of ____ ideas that were very powerful in American management. The first
 7 focusing entirely on the dining-room dimension of language. Thus the ____ linguistic point of view ends up indistinguishable from that of
 8 contemporaneity, leading one paradigm rather than another to become the ____ model. Once a new paradigm takes hold there is an
 9 you think you have made an important discovery counter to ____ opinion? Over lunch a number of years ago, one senior
 10 already contained in existing human rights instruments. Capitalizing on the ____ sentiment at the Copenhagen meeting that the time was ripe

All the lines above miss the same word. Which of the following should be that word?

A current

B indicative

C prevailing

Question 20

1 and the subsequent controversial rise and fall of Rajneeshpuram helped _____ a climate of public hostility (and ultimately judicial hostility) to
 2 weeds is therefore linked, in part, to early efforts to _____ land use on the ranchettes. The market explanation does not
 3 of forests. Along the other, however, farmers and foresters will _____ production and shrink their footprint. Forests will spread anew to
 4 in women's and men's labor and livelihoods as cotton production _____ ed are read against processes of cultural production, particularly the production
 5 1966, which brought the Baath party to power, Egyptian-Syrian rivalry _____ ed, and support for the Palestinian cause was, as usual, one
 6 along bajos are a particularly important ancient Maya technique for _____ ing agriculture because they indicate that bajos were used for crops.
 7 are needed for both sides. The U. S. government is now _____ ing direct bilateral outreach efforts with both countries that have SWFS
 8 thereby reducing the number of actively breeding females and thus _____ ing male-male competition for the few females that do manage to
 9 to taboo preserves human resources in a period when work _____ s and labor demands are increasingly placed on people who already
 10 sin of terrorism may be that it apes or even _____ s the worst aspects of the oppressive regimes from which the

All the lines above miss the same word. Which of the following should be that word?

- A** intensify(/i/ie) **B** persecut(e) **D** strengthen

Question 21

1 word value. Three definitions of intrinsic can be used to _____ an investigative framework for approximating intrinsic values and can be
 2 verbal and nonverbal communication skills and to learn how to _____ his or her spare time and to take an active
 3 with did not prove controversial. More problematic was how to _____ the government, how to divide power between the center and
 4 appropriate level of cognitive conflict is ensured. Scenarios could be _____ ed and designed to confront them with new knowledge that would
 5 optimal zone of learning, the group process should be carefully _____ ed and nurtured to ensure effective and efficient group work. Knowledge
 6 groups or races as roommates. Service learning activities could be _____ ed to enhance students' opportunities to work closely for extended periods
 7 The Scaffolded Reading Experience offers an array of strategies for _____ ing and enhancing the reading of historical documents and texts. In
 8 to continue to improve the properties of materials by artificially _____ ing them on nanometer length scales and our ability to develop
 9 epistolary confirmation of this idea. The picture itself quite literally _____ s a site for the I: in it Monet's place is
 10 _____ out of us a specific set of responses but also _____ s those responses. Working our way through a narrative results in

All the lines above miss the same word. Which of the following should be that word?

A arrang(e)

B found

C integrat(e)

D structur(e)

Question 22

1 Asia is clearly more critical for China because of its _____ and strategic relevance. However China's economic diplomacy in the region
2 alone. Authorities such as Vowell explicitly bemoaned the close spatial _____ between aboriginal and mixed-race populations. Halfbreeds, he and
3 others alleged,
4 distinct polyploid genotypes of independent origin that are in close _____ can cross with one another, leading to even more genetic
5 other egregious acts of so-called ethnic cleansing. Because of the _____ of the north and south Kivu provinces (the epicenters of
6 Stadium if they bore increased costs because of (a) the _____ of their homes to the stadium, (b) homes located behind
7 and conquer plan of the center. Bang Chan reflects this. Its _____ to Bangkok also meant that as the center reorganized its
8 credit from any of these people, no matter their actual _____ to his informal personal trade network, which also includes Korean,
9 requires disclosing the ways in which the mutual dependencies and _____ s created by globalization make it increasingly impossible to designate some
10 within a single mesocosm in a way that retains realistic _____ s, interactions, and relationships. This has led some researchers to design
deeper investigation into their moral characteristics and criminal impulses.
Criminality, _____, and the legibility of race within colonial and postcolonial theory

All the lines above miss the same word. Which of the following should be that word?

- A distance
- B immediacy(/ie)
- C proximity(/ie)
- D region

Question 23

1 of states. It has become fashionable among political scientists to _____ a causal link between democracy and peace, extrapolating from the
 2 how she uses her body, as Jardine would say, to _____ an image different from that of the traditional image of
 3 loyalty to existing countries, whatever their failings. Certainly, those who _____ continued boundary stability are correct to argue that, at present,
 4 This puts the authors in a paradox: how can they _____ students as clients, with the power of choice, and also
 5 value individual expression less than low-context cultures, some researchers have _____ ed a relationship between culture and C.A. Gudykunst and Ting-Toomey (1988),
 6 perspective) with a specific cultural group. Hence, if the propositions _____ ed by the theory are observed, then it is assumed that
 7 political violence and another led by the Partido Popular, which _____ ed that the laying down of arms by ETA would actually
 8 on the relationship between faces and fictional discourse. Though naively _____ ing a distinction between the inner and the outer may ultimately
 9 accelerated in the last years of the British Mandate. Second, _____ ing any sort of continuous American support for the Zionist project
 10 from an articulation of the booster discourse. Specifically, the theory _____ s a set of common strategies that boosters follow to "grow"

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|-------------------|------------------|----------------|------------------|
| A argue(e) | B explain | C posit | D suggest |
|-------------------|------------------|----------------|------------------|

Question 24

1 although far fewer than the stories they must organize, still a fairly ample range. For example, creation stories may be
 2 move toward an electronic, worldwide market. That market will probably all time zones and ignore all national boundaries. It will
 3 and children's literature. Brenner's major translations of creative works ostensibly the years 1930 to 1944. 14 Three are novels by Mexican
 4 Marshall went on to enjoy a major league career that ed 14 years. He set several pitching records in the 1970s
 5 in 1796 to defend the fledgling nation, Point Peter's role ed a fractious 25 years of early American history. It was
 6 deep commitment to embodied inquiry. Two teachers, whose lives have ed the century, give a particular flavor of the kind of
 7 Mexico; the Very Long Baseline Array, with ten eighty-two-foot dishes ing 5,000 miles from Hawai'i to the Virgin Islands; and the
 8 with high-incidence disabilities at five points over a 12-year period ing from adolescence through early adulthood. Whereas results contribute to our
 9 Mississippi Delta, as defined by the editors of this journal, s 210 counties in seven states. This makes it difficult to
 10 may help lower the blood pressure. Practice nurse involvement potentially s a whole range of interventions from promoting healthy lifestyle, through

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|--------------------|------------------|------------------|---------------------|
| A includ(e) | B prolong | C span(n) | D stretch(e) |
|--------------------|------------------|------------------|---------------------|

Question 25

1 but throughout eastern Kansas and even western Missouri. Historians often _____ events after 1861, though technically a part of the region's
 2 1995). In the prior study, factor analysis was used to _____ environmental-ethics scale items. We selected scale items from each factor
 3 if the Gaussians are sufficiently well separated then one can usually _____ together all the samples originating from the same Gaussian. As
 4 learn vibrato and some will not -- it is necessary to _____ your students accordingly. In addition, because students' abilities to learn
 5 secondary purpose was to determine differences in perceptions among teachers _____ ed according to the following background variables: (1) degree level, (2)
 6 with the Indian school occupying the top of the hill; _____ ed around the school, on the hillside, were representatives of various
 7 all goods in an individual household, though goods were frequently _____ ed in categories or only given the barest description by valuers.
 8 noted the sources of their information and reactions. Sources were _____ ed in the following six categories: friends, parents, teachers, media, books
 9 analysis encompasses a number of different algorithms and methods for _____ ing objects of similar kind into respective categories to develop taxonomies.
 10 advantages to using this approach outweigh disadvantages, the scheme often _____ s together individual occupations into broader categories for coding purposes. For

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|------------|----------|---------|----------|
| A divid(e) | B gather | C group | D select |
|------------|----------|---------|----------|

Question 27

1 the violent break-up of Yugoslavia. It is not unreasonable to a heightened sense of territoriality in these countries, translating into
 2 beliefs, hopes, fears, prejudices, and knowledge will lead us to different authorial and narrative audiences as we infer these positions
 3 mid adolescence (test for interaction: $p < 0.001$). # Alternatively, one could also that adolescents with high consumption of substances and behavioral problems
 4 of AT. After a careful reading of the literature, they ed and found three distinct clusters of behaviors identified as AT:
 5 (Beck & Beck-Gernsheim, 2002; Giddens, 1992). In other words, we ed teenage sexual behavior would be less influenced by membership in
 6 found that married people were happier than single people and ed that marriage may draw persons who are already more satisfied
 7 role modeling of maternal smoking behavior. Notably, parental smoking is ed to demonstrate pro-smoking norms and solidify pro-smoking attitudes. 17 18

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|----------------------|------------------------|------------------|--------------------|
| A formulat(e) | B hypothesiz(e) | C suggest | D verify(i) |
|----------------------|------------------------|------------------|--------------------|

Question 28

1 commitments in their lives such as family, employment, and other _____ activities and hobbies that capture their interest. The college players
 2 frequent, use. If their use of these drugs is often _____ and associated with leisure activities, then it is likely that
 3 appear to yield different results from comparisons between competitive and _____ athletes. # Furthermore, an examination of specific categories of superstitious
 behaviors
 4 most marinas. # A large number of respondents also commented that _____ boaters are regarded as "fat cats" and are being "picked
 5 defined polydrug use as the use of 3 or more _____ drugs in the past 90 days (meth-amphetamine, cocaine, crack, MDMA
 6 on how well the library was meeting users' educational and _____ needs so improvements could be made. They agreed to complete
 7 to the United States, where most bicycles sold are for _____ use, most of the 105 million bicycles sold each year worldwide

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|-----------------|----------------------|--------------------|-----------------------|
| A active | B distracting | C enjoyable | D recreational |
|-----------------|----------------------|--------------------|-----------------------|

Question 29

1 However, the droughts in 1969-73 and 1984 have considerably diminished _____ and pastures, thereby weakening some parental, elderly, and chiefly leverage.
2 island laboratories, where they search for new weapons against this _____ disease that threatens no consumer's health, but every consumer's pocketbook.
3 available in the state food markets, and anyone caught slaughtering _____ illegally may be imprisoned for up to four years, but
4 for the spread of sleeping sickness. 4 The restocking of _____ in Teso has been identified as a central activity for
5 the local lumber yard, they could contract to ship their _____ to South St. Paul, and they could deal with local grain
6 injure his family, and then escape with some of his _____. When they return, Farooq has gathered friends to help him
7 acquired through purchased conservation easements while ranchers continue to raise _____. This approach has become increasingly popular, especially among NGOs such

All the lines above miss the same word. Which of the following should be that word?

A breeding

B creature

C livestock

D stray

Question 30

1 their fears -- the fear that computers will mean their own and join to help encourage an area that could strengthen
 2 the number of back disorders. # However, injury reduction and eventual can only be accomplished through the joint efforts of employees
 3 fewer providers, such as E.I. du Pont de Nemours and Co.'s highly publicized 1992 of 90 percent of its outside firms that left just
 4 definition and clarification of hazing to ensure proper enforcement and of all negative behaviors. # In the fall of 2007, a
 5 block grant in the 1980s when the Reagan administration emphasized of governmental agencies and regulations. Moreover, Flenning was the first
 6 authority common to all treaty committees. The Committee on the of Racial Discrimination has benefitted from the international system's strong
 7 sword and some other cool weapons especially designed for vampire , and the fight proceeds, more or less, as one might

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|---------------------|----------------------|------------------|------------------|
| A completion | B elimination | C process | D removal |
|---------------------|----------------------|------------------|------------------|

Appendix 4. ConCloze 6 Items

Question 1

1 which the categories of Islamic anthropology and analysis are to be operated and investigated,
2 other sexual and drug use behaviors, and and demographic characteristics, we aimed to further
3 brightest and best students a sense of and intellectual responsibility. It is able to
4 of the Christian community and the subsequent and political expectations which were raised. His
5 of the three subscales, Personal Behavior Difficulty, Behavior Difficulty, and Emotional Behavior Difficulty, indicated
6 context of male family members' encouragement, women's pressure, and the positive reinforcement of praise
7 human society shaped by culture. In fact, scientists and historians are much more prone

All the lines above miss the same word. Which of the following should be that word?

A common	B economic	C important	D social
----------	------------	-------------	----------

Question 2

1 Opportunities exist for conflict prevention and for _____ agencies and organizations to assist in the
 2 landscape strewn with duplication of facilities: two _____ airports, numerous performing-arts theaters, three seaport authorities,
 3 _____ by art critics and both domestic and _____ audiences. The exception to the rule is
 4 evangelicalism in the United States, the uncertain postwar _____ climate characterized by the onset of the
 5 popular with the press and in the _____ community. Its popularity with the other ethnic
 6 onset of the Depression in 1929, as _____ financial markets progressively seized up, bankers stopped
 7 science academies, says Karim Ahmed, director of the _____ program of the National Council for Science

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|-------------------|------------------------|------------------------|-----------------|
| A cultural | B environmental | C international | D likely |
|-------------------|------------------------|------------------------|-----------------|

Question 3

1 at their own rate with students of ability. Teachers can provide more appropriate materials
 2 maternal behaviors (e.g., involvement, autonomy support) are across the two domains. Thus, the purpose
 3 multiple-line cluster to be a single entity. approach is used to determine the number
 4 two probes into this relationship yield remarkably conclusions concerning the mutual relevance and influence
 5 to someone who was involved or shared experiences is always beneficial. In a hospital
 6 can be described by a linear function, to FTT, which shows a significant reduction
 7 of electron tunneling in field ionization is to the vertical sensitivity of electron tunneling

All the lines above miss the same word. Which of the following should be that word?

A individual **B** positive **C** similar **D** various

Question 4

1 mandates are established and coupled with the _____ accountability and interventions, local schools are energized
 2 art to reform society, often without the _____ attention to the quality of these artistic
 3 and tools to respond to crises are _____ but insufficient without a firm decision and
 4 as a question about the constitutive conditions _____ for character as a possible individual to
 5 universal participation and self-sacrifice becomes perceptively less _____ for the nation's survival. Additionally, since the
 6 the business. Competency in managing change: Having _____ knowledge, skills and attitudes in order to
 7 not have access to the very sources _____ to achieve this meaning in their lives.

All the lines above miss the same word. Which of the following should be that word?

A effective B global C necessary D present

Question 5

1 and the illusion of surprise among its adversaries. It cannot massively test these systems
2 makes the Dinka positive self-perception a significant asset for development. The Dinka now demonstrate
3 to further research, as will be a[n] comparison with samples of colon unaffected by
4 disease." Plus, the district fears distractions and danger from having an animal in the
5 may help control spasticity. Because of the effects on developing cognitive skills in children,
6 and the United States, economists began assessing the impact of this substantial policy change in
7 australopithecines the benefit of the doubt as tool-users; if competition forced their extinction, their

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|--------------|-------------|------------|------------|
| A additional | B potential | C previous | D standard |
|--------------|-------------|------------|------------|

Question 6

1 and therefore it is his _____ childhood experience that is to
 2 comparative framework and not other _____ contexts such as Rwanda or
 3 her rape but the enduring _____ effects of it on her
 4 has been fully experienced, the _____ event moves from the present
 5 around two points: first, that _____ events have usually been those
 6 was clear that the woman's _____ experience during World War II
 7 of coping with and expressing _____ issues and experiences. The individual

All the lines above miss the same word. Which of the following should be that word?

- A** anxious **B** conditioned **C** disagreeable **D** traumatic

Question 7

Liberalization of the economies in Northeast Asia is as yet
to be a problem. Unfortunately, the problem has been given
use of lower-frequency (e.g. annual) price information would result in
State Departments of Education. The Court found that there was
or cardiac death, however, indicates no benefit. Thus, ESPRIT provides
number of qualified braille transcribers, to examine the issues of
overall health status (Mellor & Lindeman, 1998). Factors such as
jobs, but employers report that despite large numbers of applicants,
on their own. Yet the food supply may still be
to the home market. Similarly, if such diversion is still

All the lines above miss the same word. Which of the following should be that word?

A available

B insufficient

C meager

Question 9

1 resulted in a purchase or _____ action at the advertiser's site.
 2 well suited to provide the _____ baseline of AIS-selection information. As
 3 a list of choices. The _____ category may then be selected
 4 The opportunity to claim a[n] _____ identity was operationalized as winning
 5 coils that can generate the _____ magnetic flux. These methods also
 6 between the status quo and the _____ reference point of the community.
 7 with respect to current and _____ weight ($F = 4.85$, $p = .0134$). Tests of

All the lines above miss the same word. Which of the following should be that word?

A amorous **B** desired **C** gained **D** passionate

Question 10

1 of transition planning. Another area that warrants _____ attention is the role of disabilities on
2 electoral rules that overrepresent conservative minorities, and _____ dominance of military appointees in the judicial
3 created and provides guiding principles for the _____ existence and identity formation of the community.
4 advisors to formulate a blueprint for Taiwan's _____ growth and development in the next six
5 by increasing the demand for labor. The _____ increase in demand for labor would, in
6 develop strong rapport with participants to secure _____ participation despite new consent requirements that may
7 populations is just one index of the _____ resistance to the notion that this devastation

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|--------------------|-------------------|---------------------|---------------------|
| A continued | B everyday | C integrated | D underlying |
|--------------------|-------------------|---------------------|---------------------|

Question 11

1 to 100 degrees C and has a[n] _____ accuracy of 1-percent relative humidity. Circle 230
 2 concerning the spider web's double face, its _____ ambiguity – a distinction that will concern us
 3 model may be required because of the _____ and substantial links between language competence and
 4 its participants. Indeed, each tenet sustains the _____ and traditional values of sport, reinforcing the
 5 The emphasis on the primitive exposed a[n] _____ contradiction in evolving park policy. At the
 6 the small numbers involved because of the _____ difficulties of conducting research on African militaries
 7 bottom trawls while accounting for the biases _____ in the use of each type of

All the lines above miss the same word. Which of the following should be that word?

A emerging **B** explicit **C** inherent **D** linear

Question 12

1 long as he honored God his own actions were in the best interests of
 2 considerations, and finally you predict that this agent will act to further its goals
 3 being that are not particularly amenable to analysis assume nonreality. And thus the cosmos
 4 led to an attempt to understand the and moderate character of ancient political thought,
 5 2nd Circuit's holding that there was no basis for distinguishing between refusing unwanted medical
 6 by O'Donnell and Schmitter, was based on choice theory and related to the decreasing
 7 roles it creates – enhances the prospects for decision making. 39 In the realm of foreign

All the lines above miss the same word. Which of the following should be that word?

A objective

B rational

C representative

D residential

Question 13

1 early 1770s, dissenting denominations probably surpassed the _____ church in combined number of adherents. Especially
2 the primary coder, parsed the transcriptions and _____ criteria for coding based on previous literature
3 threaten recently democratized countries or even the _____ democracies of the region. In addition, the
4 publicly censured young activists for not following _____ guidelines for executions. Senior PLO spokesperson Bassam Abu Sharif
5 expectations for student teachers that reinforce the _____ norms and values of the school. Music
6 of civil conflicts. This task tests the _____ practices of peacekeeping, especially the circumstances in
7 Asian nuclear countries are additions to the _____ programs in Japan and Korea. So the

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|---------------|------------|---------|----------|
| A established | B indirect | C novel | D viable |
|---------------|------------|---------|----------|

Question 14

It then follows that there are other _____ activities for their time. It may therefore
Division I university. Secondly, selected market characteristics and _____ alternative sports were examined to find their
nationalist and anti-American sentiment. And because the _____ arguments reflected those driving the concurrent debate
allows for a complex exploration of the _____ claims of violence and art. Although Ofeyi,
country is still struggling to reconcile the _____ demands of security and humanity, principle and
conflict, it is not unnatural for various _____ ethnic factions to try to influence or
productively regarded as results of interaction among _____ forces. Thus the -er conjugation is where

All the lines above miss the same word. Which of the following should be that word?

A analytical

B autonomous

Competing

D territorial

Question 15

1 whereas the other does not. What is _____ about this incident is that the person
 2 to marriage has always been characterized by _____ adjustments. Nearly every society marks the onset
 3 simulation and modeling of multiphase flows. Particularly _____ are his contributions to the modeling of
 4 also because of lack of knowledge. Another _____ barrier is the small number of nurses
 5 (Stahl, 2002). Thus, the Archive data are _____ because they were collected within a public
 6 the exception. For example, a recent study, _____ for its scope and care in sample
 7 individual college level, two data compilers were _____ in their efforts to collect reliable data.

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|-----------------------|--------------------|---------------------|---------------------|
| A attributable | B disparate | C noteworthy | D sequential |
|-----------------------|--------------------|---------------------|---------------------|

Question 16

1 political action. And here Action Française becomes the _____ ally in the containment of Modernism, doctrinal
 2 persons and objects had to have "a[n] _____ and obligatory liaison," as he wrote Pelloquet,
 3 commonly viewed as such, metaphors are a[n] _____ component of science, and should not be
 4 _____ such a spirit of dialogue is a[n] _____ condition for the restoration of communio in
 5 of human rights and democratic institutions as _____ elements of a socialist society. They felt
 6 too, is often limited. Local chiefs are _____ in resolving low-level conflicts which might otherwise
 7 "to know how to write is a[n] _____ qualification for finding a proper job." The

All the lines above miss the same word. Which of the following should be that word?

- | | | | |
|--------------------|------------------------|-------------------|------------------|
| A divergent | B indispensable | C intended | D nominal |
|--------------------|------------------------|-------------------|------------------|

Question 17

1 in fact was as widely diverse and _____ along political, socioeconomic, and cultural lines as
 2 rights, this means that management rights are _____ among different and independent groups. The problem
 3 today is that they have become more _____ and disconnected than they were in the
 4 and non-Communists was transformed into a more _____ and interdependent world of competing multipolar centers
 5 have several different father figures entangled in _____ and proliferated representations. From the perspective of
 6 Operation Iraqi Freedom. Iraq is a culturally _____ country with three major ethnic/religious groups that
 7 a forest of such size in highly _____ landscapes. In such a scenario, it may

All the lines above miss the same word. Which of the following should be that word?

A contingent

B detrimental

C fragmented

D persuasive

Question 18

1 analytical and empirical work suggests that the _____ additional costs of meeting environmental requirements in
 2 than did their parents, both at the _____ and individual levels. The degree of change
 3 competition due to his estimate of the _____ annual output of the Chinese laundries in
 4 rather than as individual level phenomena. Using _____ data from thirty-four army companies, we found
 5 of work were Keynesian economies, which sustained _____ demand and kept the highly productive mass-production
 6 to the parochial, from controlling and determining _____ economic outcomes to such diverse concerns as
 7 As Panel C shows, however, medium and heavy _____ function controls have a very dramatic impact

All the lines above miss the same word. Which of the following should be that word?

A aggregate

B conspicuous

C observable

D paramount

Question 19

1 bid competitively for government contracts, or the ____ admission of black students to the state
 2 Ottawa, he declared his findings to be ____ and incomplete. In some instances, as will
 3 from a law school not recommended for ____ approval by the Committee, the Consultant shall
 4 prestige of antiquity on a recent and ____ arrangement. An imagined community requires a myth
 5 as they engage the world market. My ____ conclusion, however, is this, that American executives,
 6 III. This was accompanied by bradycardia. A[n] ____ diagnosis of inferior wall MI was formulated
 7 declared that he would support the said ____ government. Now to avoid any collision of

All the lines above miss the same word. Which of the following should be that word?

A concerted

B interdependent

C literate

D provisional

Question 20

1 in Ghana, water resource depletion in Morocco, agriculture in China, and wildlife management in
 2 midst of plenty is economically inefficient, politically and morally indefensible." In our own day
 3 between peoples made racial hybridization or amalgamation and that the long-term infertility of racial
 4 even of one single dollar, has become and will remain so as long as
 5 the long run by heaving to this approach that we have stumbled into in
 6 of reorganizing a society currently built on practices and principles. The "unite and conquer"
 7 debt at a historically unprecedented and ultimately rate. The Congressional Budget Office projects that

All the lines above miss the same word. Which of the following should be that word?

A directed

B networked

C unsustainable

D well-developed

Question 21

1 support from a growing proportion of weakly _____ and independent voters. Candidates de-emphasizing party labels
 2 and the Nimiipuu organized themselves into linguistically _____ bands, interspersed throughout the region by correlating
 3 the fall of 1987, one of the _____ companies of Japan Railways Group, Tokai, began
 4 calories per day provided to allied or _____ families and tribes versus those supporting radical
 5 was an authoritarian regime: a collection of _____ identities ruled by Serbia and disciplined by
 6 was in the United Nations and its _____ institutions. Despite the fact that Security Council
 7 were a small college, a large religiously _____ university, a large predominately minority university, and

All the lines above miss the same word. Which of the following should be that word?

A affiliated **B** centrifugal **C** indiscriminate **D** well-documented

Question 22

1 is, a voter might not know anything about a member's record but still vote
 2 in particular. In its universe the most and problematic act is writing, on which
 3 deadline. The government would avoid, however, the damages that may result if reactors are
 4 pull-out situation? Finally, what are the social effects of using local norms in the
 5 not see clearly enough that the most factor in the establishment of television in
 6 witnessed one of the most sustained and periods of school reform in our nation's
 7 conduct). For example, in the ATE program, value could be provided in terms of

All the lines above miss the same word. Which of the following should be that word?

A consequential

B imposed

C proportionate

D subsidiary

Question 23

1 his bald head and gray pajamas. The adjustments demanded by the director then involve
 2 moderation in the Reagan foreign policies were and hesitant. Administration hard-liners insisted that the
 3 to make up for weakness of the approach commonly used in education. The wholetheme
 4 if necessary. Unlike most practitioners of the approach, however, Stoffle and his associates have
 5 the nineteenth-century novel was recovered only in fashion during the post-civil war period, the
 6 supporters, the government adopted a strategy of reform via agreements with one or both
 7 problem is that the state-based approach is , such that early-enacting states must move without

All the lines above miss the same word. Which of the following should be that word?

A consonant

B mitigating

C paternalistic

D piecemeal

Question 24

1 It also reemphasizes the importance of avoiding _____ activation of the hip flexors during quiet
 2 are in parts of Iraq, but every _____ civilian death, including those among Pashtuns across
 3 to the conversation? Are names specifically mentioned? _____ disclosure of private information can occur; persons
 4 taken to protect CA resist films from _____ exposure to trace airborne quantities of such
 5 Scientists have also had a variety of _____ impacts on the dry valleys, especially before
 6 psychological distress, information (on condition and treatment), _____ self-harm, problematic behavior, alcohol, social life, intimate
 7 themselves very close to an outbreak of _____ war, and possible escalation to nuclear war.

All the lines above miss the same word. Which of the following should be that word?

A cross-national **B** germane **C** inadvertent **D** manifold

Appendix 5. ConCloze 7 Items (Modified Constructed-response, Section 4.3.4)

Question 1

- 1 which the categories of Islamic anthropology and ____ analysis are to be operated and investigated,
- 2 other sexual and drug use behaviors, and ____ and demographic characteristics, we aimed to further
- 3 brightest and best students a sense of ____ and intellectual responsibility. It is able to
- 4 of the Christian community and the subsequent ____ and political expectations which were raised. His
- 5 of the three subscales, Personal Behavior Difficulty, ____ Behavior Difficulty, and Emotional Behavior Difficulty, indicated
- 6 context of male family members' encouragement, women's ____ pressure, and the positive reinforcement of praise
- 7 human society shaped by culture. In fact, ____ scientists and historians are much more prone

All the lines above miss the **same** word. What should be that word?

- A** common
B economic
C important
D social

Question 2

- 1 and therefore it is his _____ childhood experience that is to
 2 comparative framework and not other _____ contexts such as Rwanda or
 3 her rape but the enduring _____ effects of it on her
 4 has been fully experienced, the _____ event moves from the present
 5 around two points: first, that _____ events have usually been those
 6 was clear that the woman's _____ experience during World War II
 7 of coping with and expressing _____ issues and experiences. The individual

All the lines above miss the **same** word. What should that word be?

- A** anxious
B conditioned
C disagreeable
D traumatic

Appendix 6. Verbal Reports for Processing Analyses in ConCloze 2–4 and 7 (Sections 3.3.4 and 4.3.2)

NB: Verbal reports to Item 2 (target word: *applicable*) are analyzed in both ConCloze 2–4 (Section 3.3.3 Testing usability and Section 3.3.4 Substantive content) and ConCloze 7 (Section 4.3.2 Test-taking processes and strategies). In ConCloze 7, the reports undergo a revision, which includes adding more details to the verbalizations for greater comprehensibility when segmented and migrated to the datasheet for aggregate analyses (cf. page 228 for details). To avoid redundancy, only their reprocessed version is provided in this appendix.

ConCloze 2: Aaron on Item 2 (Target word: *applicable*)

Aaron: Er.

Researcher: The way you think. [P] I would like to know the way you think.

Aaron: Why the similar [P] type of questions occurs again? [P] Erm I think this one will be harder than the previous one because I've, until now, I have answered the fourth question and [P] according to my memory, I find that er [P] each question is harder than the previous one. So, I think this one will be the hardest of the four, four questions [the math problems and Item 1 seemed to be included].

R: It's alright. I just [P] wonder how you get to the answer. Just say whatever that you are thinking.

Aaron: Hmm, [P] erm, [P] I don't have other clear message, [IA 'clearer'] message. I think I will just read the three choices, and put the [P] three words back to the sentences. [P]

R: Uh-huh.

Aaron: [Line 1, in part, left-hand only] '[IA 'traits'] which are characterized by relative stability over time [P] and [P] [KWIC position here]' er [Line 2, in phrase, from right before the KWIC position, with KWIC inserted] 'are universally [E] universally [E] [P] [Option B] 'suitable'?' No. Hmm. [Line 3, in phrase, right before the KWIC position, with KWIC inserted] 'that require you [E] clear [E] the [P] hmm [P] [Option A] 'applicable' [P]

R: Why?

Aaron: Hmm, w well I read the

R: What what word do you pay attention to?

Aaron: 'Laws' [from Line 3]. I think [P]

R: 'Laws'?

Aaron: Yeah, yeah, er in the the first sentence I think er [P] the three words are all suitable for for the first sentence, but when I read the three [P] subquestion [i.e., Line 3] I think 'laws' [word from Line 3] are already real ones and the [IA] I also I [IA [Option C] 'true'] [P] to describe laws. I think it's repeatable. So, [P] erm [Option C] 'true' it's, for [E] me, [E] I think it's maybe it's not very appropriate. And then when I look at the fifth question [i.e., Line 5], have [P] 'measurement methods' [from Line 5] er in my memory, I think, 'methods', 'measurement' [from Line 5], this kind of word are usually be described by [Option A] 'applicable'. So, so, [P] my first impression is [Option A] 'applicable'.

R: Alright, thank you. The next, please.

ConCloze 2: Björn on Item 2 (Target word: *applicable*)

Björn: The same thing, again.

Researcher: Quite, yeah.

Björn: OK.

R: Just say whatever comes to your mind. The way you think.

Björn: OK. Actually, I'm just say, speak it out and try to have the [IA] find the link between the sentence and with er [Line 1] 'tr traits, which are charac characterized by relative stability over time and hmm something [i.e., KWIC position] across situations posi positive state-like capacities relatively more [P] more malleable and' er the second one is the [Line 2] 'urbanization display three research perspectives [P] abstract frameworks that are [E] universally [E] hmm [P] [KWIC position here] but at the expense of the su substantive specily [in fact 'specificity'] context-dependent frameworks that' And [Line 3] 'require the use of resources that require you to, you clear the [P] [KWIC position here] copyright laws before viewing and listening to them with your' [P] [Line 4] 'aim for a break, just for for socialist reform, a reform [KWIC position here] in the new areas opened by the revolution and its' [P] [IA] 'something like that' [IA] [P]

R: Just say. Keep saying.

Björn: And [Line 5] 'still to be resolved and n374, of 188 H-A-P-s'

R: It's OK. You, just just say, yeah, whatever comes to your mind. You don't have to read everything.

Björn: OK.

R: Just which words you are reading, yeah.

Björn: OK.

R: The way you think. I would like to know the way you think. Say it.

Björn: Hmm, [P] haha [laughing] say something. er [sigh] haha I'm embarrassed.

R: You can explain your answer, for example.

Björn: Hmm, I think I can get some hints from Sentence 4 [i.e., Line 4]. So, it's er just [Line 4, in phrase, right before the KWIC position] 'socialist reform [KWIC position here]'? But what erm, 'reform' [from Line 4] is quite er what's the [Line 4 again immediately, in phrase, right before the KWIC position] 'reform [IA] [KWIC position here]' what [IA]. All the questions is from same erm same article?

R: No.

Björn: No? OK.

R: They are from different places.

Björn: Different places. Hmm. [P] what's what's [IA] [Line 6, in part, left-hand only] 'While this solution may not be [KWIC position here]' er [Line 6 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'may not be er [Option B] 'suitable''? [P]

R: Go on.

Björn: So, I think that answer [P] may be choice A or B? [Option A] 'applicable' or [Option B] 'suitable'.

R: Uh-huh.

Björn: Becoz er [P]

R: Why?

Björn: But's er, but which one will be the most [P] the most suitable one for all the ques[tion] er for all the sentences. [P] [IA] So, erm [P] from the last sentence [i.e., Line 7] I think is 'suitable' [Option B] 'suitable'? Answer B? [Line 7] 'Court held that so long as law is [E] generally [E] [Option B] 'suitable'? [Option B] 'suitable'. [Line 7 continued] 'and not discriminatory in being specifically aimed at religion' [IA] [Line 7 again immediately, in part, left-hand only, with KWIC inserted] 'Court held [IA] so long is [E] generally [E] [KWIC position here]' [Line 7 again immediately, in word, right-before the KWIC position, with KWIC inserted] 'generally [Option B] 'suitable''? [Line 7 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'law is generally [Option A] 'ap applicable'', I mean.

R: It's OK.

Björn: [Option A] 'applicable'?

R: OK, the next question please.

ConCloze 2: Claire on Item 2 (Target word: *applicable*)

Claire: Every every [P] er every number is er do not

Researcher: Not from the same place.

Claire: OK.

R: They are from different places.

Claire: OK, OK. [P] same, same, the same item.

R: Yeah.

Claire: OK.

R: The same form of item.

Claire: [Line 1] '[IA 'traits, which'] are [IA '[E] characterized [E]']

R: Say it [P] clearly.

Claire: [Line 1 continued] 'by relative er stability over [E] time [E] and [P] [KWIC position here] over time? And er blank [i.e., KIWC position] [P] across [E] situations? [E] positive state-like capacities are [P] psst hmm [P] relatively more [E] malleable [E] and' [P] [Line 2] 'urbanization display [IA 'three'] research perspectives

R: Say it clearly, loudly.

Claire: [Line 2 continued] 'abstract frameworks that' [P] er I think I need to pass the yeah [P] next

R: Uh-huh. OK. Say it

Claire: [Line 2 continued] 'are [E] universally [E] [KWIC position here] but [P] at the expense of [P] psst substantive specificity [IA 'context-dependent'] frameworks that' [P] [Line 3, in part, left-hand only] '[IA 'use of resources'] that require you [P] [E] clear the [E] [KWIC position here]' [gasp] [Line 3 again immediately, in part, from right before the KWIC position] '[E] clear the [E] [P] [KWIC position here] copyright laws [P]'

R: Say it, whatever you are thinking.

Claire: [Line 3 continued] 'before viewing and listening [IA 'to'] them' [P] I think it's the erm [P] hmm [P]

R: Say it.

Claire: Number A or number B?

R: Uh-huh.

Claire: [Option A] 'applicable' or [Option B] 'suitable'?

R: Uh-huh, [P] OK. Why?

Claire: I think er the 'copyright laws' [from Line 3]? [P] Yeah, because you don't [IA 'noun nears'] 'the copy [IA 'laws']' [from Line 3] [P] [E] 'copyright laws' [from Line 3]

R: Uh-huh.

Claire: So [P] I think er I if I put [P] 'true copyright laws' I think er it doesn't [IA] make sense.

R: Uh-huh.

Claire: Anyway, I'll read more [P]

R: Uh-huh. Good, good. Say it loudly, clearly.

Claire: [Line 4, in part, left-hand only] 'aim [IA 'for a'] break [P] just for socialist reform, a reform? [P] [KWIC position here]' [Line 4 again immediately, in part, from right before the KWIC position] '[E] reform [E] hmm [P] [KWIC position here] in the new [E] areas [E] [P] opened by the [P] revolution and its' [Line 5] 'still to [P] [six words not verbalized] have er [P] [KWIC position here] ['measurement methods' not verbalized]. '43 have likely methods [IA] ['Eleven have' not verbalized] potential or' [P] [Line 6, in part, left-hand only] 'efficient [P] ['and biologically' not verbalized] appropriate er [P] hmm ['While this solution may' not verbalized] not be [P] [KWIC position here]' [Line 6 again immediately, in part, from right before the KWIC position] 'not hmm [P] [KWIC position here] to all common property resources [IA] for those cases [P] which may' [P] [IA] [Line 7, in part, from right before the KWIC position] 'law is [E] generally [E] [P] [KWIC position here] and not discriminatory' [P] [gasp] er erm I felt confused, [Option A] 'applicable' or er [Option B] 'suitable'.

R: Uh-huh.

Claire: Hmm [Option A] 'applicable' hmm

R: Whatever you are thinking, reading, just say it.

Claire: [P] [IA 'appropriate'] [from Line 6] hmm [P] I think [P] er the hint word is [E] 'appropriate' [from Line 6] [E]? [P]

R: Uh-huh.

Claire: Er word in number 6 [i.e., Line 6]?

R: Uh-huh.

Claire: [P] Yeah. So, I think the [Option B] 'suitable' [P] [gasp] ah confused

R: It's OK. No worry.

Claire: [Line 6 again immediately, in phrase, right before the KWIC position] 'may not be [P] [KWIC position here]' [Line 6 again immediately, in phrase, right before the KWIC position] 'may [KWIC position here]' [P] [Line 6 again immediately, in part, but started a little earlier, with KWIC inserted] 'While this solution may not [E] be [E] [P] [Option B] 'suitable'? To all [E] common [E] property resource.' [P] or [Line 6 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'this solution may not be [Option A] applicable' [P] 'to all common [IA 'property resources'] for those cases which [E] may [E]' [P] [Line 7, in word, right before the KWIC position, with KWIC inserted] 'generally [IA [Option A] 'applicable']' [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] '[E] generally [E] hmm [P] [KWIC position here]' [Line 7 again immediately, in word, right before the KWIC position] '[IA 'law']' [Option A] 'applicable', [Option B] 'suitable' [P] hmm [gasp] ah

R: Say it.

Claire: [P] I'll I'll go back to the first two sentence.

R: Uh-huh.

Claire: [Line 1, in phrase, right before the KWIC position, with KWIC inserted] '[IA] stability over time and [P] [Option A] 'applicable'' [P] I think it's number 1 [Option A] 'applicable' [P] because it's er 'across situations' [from Line 1]

R: Uh-huh.

Claire: Yeah, it's more reasonable. [Option A] 'Applicable' is more reasonable word. Yeah, so [Option A] 'applicable', yeah.

R: OK.

Claire: Really difficulty.

ConCloze 2: Dakota on Item 2 (Target word: *applicable*)

Researcher: Quite.

Dakota: Oh, my goodness. We [IA], I don't think we got er we have, oh no, it is just the survey, right? Here.

R: It is the prototype. Let's say, the pilot version.

Dakota: Yeah.

R: Which is still under development.

Dakota: [Line 1, in part, left-hand only] '[traits' not verbalized] which are charac characterized by [P] relevant relative [actually just 'relative', without 'relevant'] stability over time and [P] [KWIC position here]' [Line 1 again immediately, in phrase, right before the KWIC position] 'over time and [P] [KWIC position here]'

R: Say it, whatever you are thinking.

Dakota: [Line 1 again immediately, from right before the KWIC position, with KWIC inserted] 'over time and the [Option B] 'suitable' across situations'. [Line 1 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'and [Option A] 'applica[ble] [P] [Option A] '[E] appli [E] [P] applicable'' [P] ha?

R: Alright.

Dakota: [Line 1 continued] '[Option A] 'Appli applicable applicable' across' [P] [Line 1 again immediately, in phrase, from right before the KWIC position] 'time ['and [KWIC position here] across situations' not verbalized] positive' erm [Line 1 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'sta stability over [E] time [E] and er [Option B] 'suitable' across situations' [P] and [Line 2, in phrase, right before the KWIC position] 'the frameworks that are [KWIC position here]' [P] [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'universally [P] [Option B] 'suitable' but at the [P] [IA] expense of' the ok? [Line 3, in part, with KWIC inserted] 'require the use of er resources that require you cl clear the [P] [Option A] 'ap applicable applica' ['copyright' skipped] laws before' ok? [Line 4, in phrase, from right before the KWIC position] 'a reform [KWIC position here]' [P] [Line 5, in phrase, from right before the KWIC position, with KWIC inserted] 'one three four have [Option A] 'applicable' er measurement methods' [Option B] 'sh suitable' [Line 5 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'have [Option B] 'suitable' measurement methods' [P] [Line 3, in phrase, from right before the KWIC position, with KWIC inserted] 'have [E] clear [E] the [Option B] 'suitable' copy [IA 'can't be'] [Line 3 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option B] 'suitable' copyright [P] laws' [Line 3 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'the [Option B] 'suitable' copyright laws'? No. Psst er [Line 6, in part, from right before the KWIC position, with KWIC inserted] 'this solution [E] may not be [E] [Option A] 'applicable' to all common [P] [IA 'property'] resources' [Line 7, in part, from right before the KWIC position] 'law is [E] generally [E] [P] [KWIC position here] not discriminatory in being' [P] [Line 7 again immediately, in part, wider part] 'so long as a law is generally [KWIC position here]' [P] I think A. [P] But A [Option A 'applicable'] and B [Option B 'suitable'] are similar haha [laughing]

R: Uh-huh.

Dakota: Er [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] 'over time and [KWIC position here] the [IA '[Option A] 'applicable' across situations']' [P] Yeah I think maybe A [Option A 'applicable'].

R: Which one, 'over time' [from Line 1]? From Line er Line 1, right? Uh-huh. Anything else that helps you to get to the answer?

Dakota: [Line 2, in phrase, right before the KWIC position] 'are [E] universally [E] [KWIC position here]'

R: Any key word? Or any words that you think are

Dakota: Yeah, yeah, I just the I just the put the words here if the if that makes sense to the

R: What do you mean, 'make sense'?

Dakota: 'Make sense' is is means like er you read like er sentence like er they did er they [IA 'mean er'] something. They did er say something. But actually these two words [Option A 'applicable and Option B 'suitable'] two words quite close, I mean. [P]

R: It's OK.

Dakota: Yeah, I think.

R: I want to know the way you think. No worries.

Dakota: I just I just choose this one [Option A 'applicable']

R: OK. Alright. OK.

ConCloze 2: Esther on Item 2 (Target word: *applicable*)

Esther: So, same kind of question?

Researcher: Yes.

Esther: Right. And the words [Option A] 'applicable', [Option B] 'suitable', and [Option C] 'true' [P] again, I think I look for the [IA 'sentence in which'] [P] yeah, [IA 'where'] the beginning is there. [P] [Line 6, in part, with KWIC inserted] 'While this solution may not be [Option A] 'applicable' [P] to all common property resources for those cases which may' [P] Could be. [P] Now I see over here [P]

R: Why so?

Esther: Er, [P] [Option B] 'suitable' may not be [P] ers like [P] suitable to use here because [Line 6 again immediately, in part, with KWIC inserted] 'While this situation may not be [Option B] 'suitable' [P]

R: Uh-huh.

Esther: To well actually [Line 6 continued] '[IA] [KWIC position here] to all common property resources' But I think [Option A] 'applicable' [P] sounds better over here?

R: For Line 6, eh?

Esther: [Line 6 again immediately, in part, with KWIC inserted] 'While this solution [E] may [E] not be [Option A] 'appli[cable]' er [Option C] 'true' to all common property resources.' [P] But to me, [Option A] 'applicable' is [P] sounding better.

R: What about other lines? Do they help?

Esther: [Line 4, in part] 'aim for a break just for socialist reform, a reform [P] dash [i.e., KWIC position] in the new areas opened by revolution' erm [P] again because the beginning of the sentence is not there, so I can only make a guess. But [P] [Option A] 'applicable' can be even used over here?

R: Can or cannot?

Esther: Can, I think.

R: Uh, Line 4. What about other lines?

Esther: Because I don't actually [E] know [E] what this [P] what message this sentence is conveying.

R: Uh.

Esther: So, [P] [Line 4 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'a reform [Option C] [E] 'true' [E] in the new areas' [P]

R: Uh-huh. OK.

Esther: But erm [P] OK, [Line 7, with KWIC inserted] ‘Court held that so long as a law is generally [P] [Option C] ‘true’ [P] and not discriminatory in being specifically aimed at religion’ er [P] haha [laughing] I got the there’s nothing more after the, so it’s become too difficult to understand. [P]

R: Uh-huh. So,

Esther: [Line 7 again immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘Law is [‘generally’ not verbalized] [Option A] ‘applicable’ even erm [Option A] ‘applicable’ can be used with ‘the law’ [from Line 7], so [P]

R: Uh.

Esther: So, if I look at this one, [Line 1, in part] ‘traits, which are characterized by relative stability over time [P] and [KWIC position here] across situations’ [P] [Line 1 again immediately, in part, with KWIC inserted] ‘which are characterized by relative stability [P] over time [P] and [Option A] ‘applicable’ across situations [P] positives state-like capacities are relatively more malle[able]’ [P] So, it makes sense over here. [P]

R: Which word? Which word?

Esther: Same, [Option A] ‘applicable.’

R: [Option A] ‘applicable’. OK.

Esther: Now, something somethings [Line 2, in phrase, right before the KWIC position, with KWIC inserted] ‘[the given subject ignored] are universally [P] [Option C] ‘true’’, you can say.

R: Er, this is Line 2; it’s too long.

Esther: Oh, OK, OK. [Line 2 again immediately, in part, with KWIC inserted] ‘urbanization [E] display [E] three research perspectives[:] abstract frameworks that are [E] universally [E] [P] [Option C] ‘true’ but at the expense of substantive substantive er specific[ity]’ but over here, it’s also like because er [P] [Option C] ‘true’, [Option B] ‘suitable’, and [Option A] ‘applicable’, these words are kind of, you know, having more or less the same meaning or like closely [P] related, you know.

R: Yeah.

Esther: Related. They are not opposite to each other. So, that’s why unless you know the whole sentence er you know the theme or you know er what message this sentence is conveying, it’s difficult to choose one of these.

R: But why do you choose [Option A] ‘applicable’ over [Option B] ‘suitable’, or over [Option C] ‘true’, if they are really close?

Esther: Because it’s [P] kind of [P] makes more sense?

R: Make more sense? OK, Yeah. It’s OK. Anything else that you would like to say?

Esther: Erm hmm [P] And over here I think [Line 3, in phrase, right before the KWIC position] ‘the [KWIC position here] law laws’ are more

R: ‘Law’ [from Line 3], uh-huh. So, the key word for you is ‘law’?

Esther: Possibly. [P]

R: Uh.

Esther: OK, let me see this one. [Line 5, in part] ‘still to be resolved [P] 188 HAPs, 134 [E] have [E] [KWIC position here] [P] measurement methods.’ [Line 5 again immediately, in phrase, from right before KWIC position, with KWIC inserted] ‘have [KWIC position here]’ [P] again you can use [Option B] ‘suitable’, ‘suitable’ measuring measurement methods’. But you can’t say [Line 5 again immediately, in phrase, right after KWIC position, with KWIC inserted] ‘[Option C] ‘true’ measurement methods’. [P] [Line 5 again immediately, in phrase, right after KWIC position, with KWIC inserted] ‘[Option A] ‘applicable’ measurement methods’, it’s also suitable word. And [Option B] ‘suitable’ is also, it can go very well here. But [Option C] ‘true’, no, doesn’t fit here. [P]

R: Alright.

Esther: It's kind of like erm the least possible I'm cancelling out, and then I'm narrowing down. That's why I, that's how I'm selecting.

R: Yeah, OK. That's good. Thank you.

ConCloze 3: Franz on Item 2 (Target word: *applicable*)

Researcher: Now, could you please continue to the second question? Say whatever you are thinking, whatever you are [P], yeah, doing in your mind.

Franz: Hmm, [P] for this time, I try to translate the choices [P] first, haha [chuckle] because it can be more easier.

R: Yeah.

Franz: It can be [E] much easier [E], sorry

R: It's ok.

Franz: For my grammar.

R: No worries.

Franz: Haha [laughing] I always detect myself. [E] I [E] do always I always do auto-correct myself. That's horrible. Trust me.

R: That sounds interesting. OK.

Franz: Trust me. [P] Erm, well, [Item instructions] 'All the lines above miss the same word. Which of the following should be that word?' Erm, [Option A] 'applicable', [Option B] 'suitable', and [Option C] 'true'. They are all different. [P] Right?

R: Yes.

Franz: So, [P] after finding their meanings, now I do take a look [P] at certain [P] sentence.

R: Yes.

Franz: [Line 1, in part] 'traits, hmm which are characterized by relative stability over time and [P] [KWIC position here] hmm? across situations'. [IA 'what's'] tut tut tut tut [= 'blah blah blah']

R: Hmm.

Franz: OK, it must be adjective.

R: Yeah.

Franz: But I guess from this second sentence [i.e., Line 2], not the first one [i.e., Line 1]. The first one is quite [P] confusing.

R: Yeah.

Franz: Right? Umm [P], yes, er, the fir first one is quite confusing. But I know it must be [P] adjective, and all answers are [E] adjectives [E].

R: Yes.

Franz: Apparently. Erm [sigh] pressurized, pressurized, ohh hahaha [laughing]

R: It's ok. Just keep saying.

Franz: Erm [P] Psst, now I'm translating. [P] actually, 'traits' [from Line 1] are something like 'characteristics'. Right? Erm, or specific characteristics, traits, traits, erm [Line 1] 'traits which are characterized by relative stability over time [P] [KWIC position here] and across situations, positive state-like specificity are relatively more [P] malleable' O, I don't know this word,

R: It's OK.

Franz: [P] 'malleable' [from Line 1] and [Line 2, in part, left-hand only] 'urbanization display three research perspectives: abstract frameworks [P] that are [E] universally [E] [KWIC position here]' Ah, interesting. [P] When I see erm the word 'research' [from Line 2] [P] my mind [P] links to the word [Option A] 'applicable' [P]

R: Uh.

Franz: Because when you do research erm it tends to [P] er to the situations that you must apply something [P]

R: Uh-huh.

Franz: But I haven't decided yet. OK?

R: OK.

Franz: Want to choose. OK? Let's continue.

R: Yes, of course.

Franz: [Chuck] [Line 2 again immediately, in part, right-hand only] '[KWIC position here] but at the expense of substantive specificity; context-dependent frameworks that' [Line 3, in part, right-hand only] 'require the use of resources that require yo you [E] clear [E] the [KWIC position here]' [Line 3 again immediately, in phrase, right before the KWIC position] 'that require you [E] clear [E] [KWIC position here] [IA]' [P] I I was wondering [P] if there's something wrong with [P] grammar of this sentence [i.e., Line 3].

R: I'm sure there is no.

Franz: There is no? OK, [Line 3 again immediately, in part, left-hand only] 'that require the use of resources that require [E] you [E] [P] [E] clear [E] [KWIC position here]' [P] [Line 3 again immediately, in phrase, right before the KWIC position] 'that require you [E] clear [E] [KWIC position here]' [P] so 'clear' [from Line 3] [IA 'can be verb']? [P] Right?

R: Perhaps.

Franz: OK, fine!

R: I am an interviewer. I cannot say much.

Franz: Haha [laughing]. [Line 3 again immediately, in phrase, from right before the KWIC position] 'clear the [KWIC position here] copyright laws', Ow! [P] When I see 'the cop the copyright laws' [from Line 3], it can be [E] [Option B] 'suitable' [E] as well.

R: Uh-huh.

Franz: [P] So, now I'm confused. Let's take a look at third sentence [i.e., Line 3]. Er, no, I mean four fourth one [i.e., Line 4]. [P] [Line 4, in part, left-hand only] 'aim to [in fact 'for a'] break, just for social socialist reform, a reform [P] hmm [KWIC position here]' [P] [Distraction from the outside] Can you hear the siren?

R: Of course.

Franz: And it means this city is not [E] quite safe [E]. Haha [laughing]

R: Er, I think it is an ambulance, perhaps.

Franz: O, really. [P] And it means you safe?

R: Perhaps, [P] it depends on interpretation.

Franz: Ok, everything [E] sounds relative [E]. [P] Erm [P] Oh, no, no, no, [P] no, no, no, hmm

R: [IA 'Let's show it.'] Clearly you are thinking a lot.

Franz: [laughing] Yes, because I was distracted by the siren, I'm sorry.

R: OK, please go ahead. Please.

Franz: [Line 4 again immediately, in part, left-hand only] 'aim for a break, just for socia socialist reform, a reform [KWIC position here]' hmm [Option B] 'suitable' sounds, doesn't sound good. I mean, if I choose [P] [Option B] 'suitable', it doesn't sound [E] good [E]. It doesn't sound proper, here, for fourth sentence, I guess. [P] er return to it. [Line 5] 'still to be resolved tut tut tut have [KWIC position here] measurement methods. 43 have likely measurement methods. Eleven have potential' [P] erm, psst, can be [Option B] 'suitable'. It can be [Option B] 'suitable' again. Haha [laughing]

R: It's OK.

Franz: For the fifth item [i.e., Line 5]! [P] [Line 6, in phrase, right before the KWIC position] 'may not to [E] be [E] [KWIC position here]'. And then this can be [Option A] 'applicable', for number 6 [i.e., Line 6]. [P] What! [P] [Line 7, in part, right-hand only] 'Court held that so long as a law is generally [P] [KWIC position here]' OK, [P] I think it's [P] Choice A [Option A 'applicable']. [P]

R: Uh-huh.

Franz: I choose Choice A, 'applicable' [Option A].

R: Could you please [P] tell me, again, the key words that help you to get to this answer?

Franz: Erm, [P] for [P] the second sentence [i.e., Line 2], it's 'research perspectives'.

R: Yeah.

Franz: For the sixth sentence [actually, Line 7], it's erm 'law[s]' and 'generally' [from Line 7].

R: Yes.

Franz: [Line 7, in word, right before the KWIC position, with KWIC inserted] 'generally [Option B] 'suitable'' [A phrase formed out of one of the options and an adverb from Line 7] doesn't sound good. But [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] 'generally [Option A] 'applicable'' [i.e., an adverb in Line 7 and Option A] sounds better. [P]

R: Alright.

Franz: So, I'm sorry. This is kind of common sense. I don't know.

R: Yeah. OK. Anything else?

Franz: No.

ConCloze 3: Gill on Item 2 (Target word: *applicable*)

Gill: The second question. [Line 1, in part, with KWIC inserted] 'traits traits which are com characterized by relative stability over [E] time [E] [P] across [P] [IA 'across'] [Option A] appli 'applicable' across', [P]

Researcher: Keep saying, yeah.

Gill: [Line 1 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'suitable' across', [Line 1 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'true' across [P]' no, [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option A] 'applicable' across, positive state-like stability' [P] [Line 1 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'a a [P] over time and [Option A] 'a a applicable''. O, sorry, a is applicable. OK, for the for the start, I'll [IA 'book'] [Option A] 'applicable'.

R: Uh-huh.

Gill: Let me move forward

Research: Yeah.

Gill: And see. [P] [Line 2, in part, left-hand only, with KWIC inserted] 'urbana urbanization display three research [P] perspectives: abstract frameworks [E] that [E] [P] are [E] universally [E] yeah [Option A] 'applicable'.'

R: Wow.

Gill: It's yeah. [Line 3, in part, left-hand only] 'It requires the use of resources' [P] [Line 3 again immediately, in part, left-hand only] 'it requires the use of resources that require you [P] [E] clear [E] [KWIC position here]' [Line 3 again immediately, in phrase, left-hand only] 'that requires that you' [P] [Line 3 again immediately, in phrase, left-hand only] 'that requires that you' [Line 3 again immediately, in phrase, left-hand only] 'that requires that [P]' a [Line 3 again immediately, in part, left-hand only, with KWIC inserted] 'requires the use of resources that require [E] you clear the [E] [P] [KWIC position here] ap appropriate' [P] no, no, no. [Line 3 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'you clear the [Option A] 'applicable' copyright laws before [P] listening to them with' ok, ok. Yeah, so [Option A] 'applicable'.

R: Uh.

Gill: Yeah, because it it requires you to do something in order to [P] because of the words 'copyright laws' [from Line 3], OK. So, [Option A] 'applicable'. Then the the fourth one. [Line 4, in part, with KWIC inserted] 'aim for a break, just for socialist reform, a [E] reform [E] [P] [Option A] 'applicable' in new areas'. Yeah, [Option A] 'applicable' [P]

R: Uh.

Gill: [Line 5, in part, with KWIC inserted] '[E] still [E] to be resolved n37 74 of 188 HAP 134 have' [P] yeah, [Option A] 'applicable' measurement methods' [P] 43 have likely methods'. [P] no [IA 'let me see'] [Line 5 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'have [Option B] 'suitable suitable' measurement methods. 43 have [P] [E] likely [E] [P] eleven have potential' [P] [IA 'it turns out'] [Option A] 'applicable' yeah 'applicable', just [Option A] 'applicable', it's [Option A] 'applicable', 'likely' [from Line 5] and 'potential' [from Line 5], ok. [Notice the word associates picked] [P] number 6 [Line 6, in part, left-hand only] 'efficient and biologically appropriate [E] While [E] this solution may not be [KWIC position here]' [P] [Line 6 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] '[E] may not be [E] s [P] [Option A] 'applicable' to common property resources', [Line 6 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'may not be [Option B] 'suitable' to all common property resources, for these cases, for those cases which may' [P] [IA] er [Option A] 'applicable'. As I said, they put [Option B] 'suitable' er just to get you confused. It's closely related to [Option A] 'applicable'.

R: Uh-huh.

Gill: [Line 7, with KWIC inserted] 'Court held that so long as [P] a law is generally [P] [Option A] 'applicable' and not discriminatory in being specifically aimed at religion' [P] yeah, it's, yeah, it's [Option A] 'appli applicable'

R: So you you said that [Option B] 'suitable' might also be possible,

Gill: Yeah, yeah, but but

R: But but why do you choose [Option A] 'applicable' [E] over [E] [Option B] 'suitable'? Why?

Gill: I I think that by the the nature of cert certain words helps me to think that [Option A] 'applicable' is suitable, you know.

R: For example?

Gill: For example, that [Line 7 again immediately, in phrase, right before the KWIC position] 'so long as a law [KWIC position here]' [P] it may be 'suitable', OK. But again, you look at the gen a general ability [the word 'generally' in Line 7?] [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] 'generally [Option A] 'applicable''

R: Uh-huh.

Gill: Because in English when there are certain words, when they come, usually you should know which words that will follow.

R: Uh-huh.

Gill: If you are a student of English, you should know that ‘generally’ it just be can’t [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] ‘generally [Option B] ‘suitable’’, but it is best to be [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] ‘generally [Option A] ‘applicable’’, yeah.

R: So, you think Line 7 helps you the most to [P]

Gill: Yeah, yeah.

R: To clear-cutly decide.

Gill: Yeah, yeah.

R: OK, please tick it.

ConCloze 3: Halle on Item 2 (Target word: *applicable*)

Halle: Again, same exercise?

Researcher: Yes.

Halle: So, erm [P] The first line is [Line 1] ‘traits, which are characterized by relative stability over time and blank [i.e., KWIC position] across situations, positive state-like capacities are rel relatively more malleable and’ [P] and the words I have are [Option A] ‘applicable’, [Option B] ‘suitable’, or [Option C] ‘[E] true [E]’? [P] er [Line 1 again immediately, in part, with KWIC inserted] ‘are characterized by relative stability over [E] time and [E] [P] [Option B] ‘suitable’ across situations’ or [Option A] ‘applicable’. I’m gonna go see the next one because, again, it is hard to decide from just the first line. So, the other lines are helping me to see which word is the correct one. [P] OK, so second line [Line 2, with KWIC inserted] ‘urbanization displays three research perspectives: abstract frameworks that are [E] universally [E] [P] [Option A] ‘applicable’? [P] but at the expense of substansive [actually ‘substantive’] specificity, context-dependent frameworks that’. I think [Option A] ‘applicable’ would work [P] or maybe [Option C] ‘true’. [P] er [P] no, I think I will go with [Option A] ‘applicable’. This is [E] hard [E] because they are quite similar this time, so it is not that hard [??] to make the difference between them.

R: Uh-huh.

Halle: I will stick with [Option A] ‘applicable’ for now. [P] third line [Line 3] ‘require the use of resources that require you [E] clear the [E] [P] blank [i.e., KWIC position] copyright [E] laws [E] before viewing and or listening to them with your’. O, in this case, [Option C] ‘true’ does not really make sense, [Line 3 again immediately, in phrase, right after the KWIC position, with KWIC inserted] ‘[Option C] ‘true’ copyright laws’? [P] [Line 3 again immediately, in phrase, right after the KWIC position, with KWIC inserted] ‘[Option B] ‘suitable’ copyright laws’ [P] might make sense. [P] But then again also, [Line 3 again immediately, in part, with KWIC inserted] ‘require the use of resources that require [P] you clear the [Option A] ‘applicable’ [P] copyright laws’. [P] It would be between [Option A] ‘applicable’ and [Option B] ‘suitable’ but for this one, I think I rather go with [Option B] ‘suitable’ maybe? I’m not sure. [Chuckling] I’m gonna go to the next one. [P] [Line 4] ‘aim for a break just for socialist reform, a [E] reform [E] [P] blank [i.e., KWIC position] [P] in the new areas opened by the revolution and its’ [Line 4 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] ‘a [E] reform [E] [KWIC position here]’ [P] I’m gonna say ‘[Option A] ‘applicable’ in the new areas’ because if it would have been [Option B] ‘suitable’, it would have been ‘[Option B] ‘suitable’ [E] for [E] [P] something’, so I’m gonna go with [Option A] ‘applicable’. [P] So the next one, [Line 5] ‘still to be resolved N three thousand seventy-four. [P] Of 188 HAPs, 134 have [E] [P] blank [i.e., KWIC position] measurement methods. 34 have likely methods. 11 have potential’. Again, I think [Option A] ‘applicable’ goes best [P] out of the three options in this case. [Line 5 again immediately, in phrase,

right after the KWIC position, with KWIC inserted] '[Option A] 'applicable' measurement methods.'

[P] Then next one, [Line 6] 'efficient and biologically appropriate. While this solution may not [E] be [E] blank [i.e., KWIC position] to all common property resources for those cases which [E] may [E]'. Well, here, definitely [Option A] 'applicable' [Line 6 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'may not be [Option A] 'applicable' [E] to [E]' [P] I think the 'to' really helps me figure out the [IA 'end-like'] prepositions, which one between [Option A] 'applicable' or [Option C] 'suitable' works best and so for [Option A] 'applicable' is the one correct. [P] and then the final line [Line 7] 'Court held that so long as a law is generally blank [i.e., KWIC position] and not discriminatory in being specifically aimed at religion [P] the'. So, again [Option A] 'applicable' works. So, I'm gonna go with [Option A] 'applicable' for all of them.

R: OK. Yeah, you can click.

ConCloze 3: Igor on Item 2 (Target word: *applicable*)

Igor: You have to cut something that is [P] 100% wrong [E] out [E], and

Researcher: Hmm.

Igor: OK, your chance to get the right answer that higher. From four choice you get three. That's mean the percent to correct is increase.

R: Uh-huh.

Igor: Erm, I don't know. OK.

R: It's OK. No problem. Just keep saying.

Igor: What is this? Second question. The same, right? This one, do the same.

R: Yeah.

Igor: Ah, this one it has [P] three [probably 'three options'], OK. [Option A] 'applicable', [Option B] 'suitable', [Option C] 'true'. OK, this meaning is no problem. But [P] [Line 1, in part, left-hand only] 'traits which are characterized by relative stability [P] [E] over time [E] [P] [KWIC position here] [IA]'

R: Keep saying.

Igor: [Line 1 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'and [P] over time and [KWIC position here]' [P]

R: Keep saying.

Igor: [Line 1 continued] '[Option A] 'applicable' across situations [P] hmm [IA]'

R: Keep saying.

Igor: Yeah, yeah. Just let me think a few minutes [P] Hmm. [IA] [Temporarily out of topic, because of a painter coming into the room]

R: Keep saying.

Igor: I know what I have to say, OK? Let me see first. Before I say, I have to think something.

R: Ouch, alright. That that, whatever you are thinking, just say it.

Igor: Now I am looking the [P] the appropriate choice to put, to put in the question. The missing word. [P]

R: Yeah.

Igor: So, the first first er [Line 1 again immediately, in part, with KWIC inserted] 'traits which are characterized by relative stability relative stability over time and [P] [Option A] 'applicable'? across situations?' Or it? It can be. Or [Line 1 again immediately, in word, right before the KWIC position, with KWIC inserted] 'and [P] [Option B] 'suitable'? [Line 1 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'and [Option B] 'suitable' across situations'? [P] It's

wrong. [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option C] 'true' across situations'? I think the last one [i.e., Option C 'true'] is not make sense.

R: Uh-huh.

Igor: And, OK. Let's go on. [IA]

R: Keep saying.

Igor: [Line 2, in part, left-hand only] 'display three research perspectives[:] abstract frameworks that are universally [P] universally? universally [KWIC position here]' [P] [IA] [P] [Line 3, in part] 'require the use of resource resource that require you [P] [E] clear [E] the [P] [KWIC position here] [P] copyright' [Line 3 again immediately, in word, right before the KWIC position, with KWIC inserted] '[Option B] 'suitable' copyright' [Line 3 again immediately, in part, right after the KWIC position, with KWIC inserted] '[Option C] 'true' copyright laws before viewing and listening' [P]

R: Keep saying.

Igor: [Line 3 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'you clear the [Option A] 'applicable' copyright law'. [P] O, this one is quite hard for me. [P] [Line 4, with KWIC inserted] 'aim [P] for a break just [P] aim for a break just for socialist reform, [P] a reform [P] [Option A] 'applicable' [P] in the [E] new [E] area opened by the revolution and its' [Line 4 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option B] 'suitable' in the new [IA 'area']' This one should be [Option B] 'suitable' haha [chuckle] psst [P] erm [Line 5, in part, left-hand only, with KWIC inserted] 'still to be re resolved [P] [Note that several words in between are not verbalized here] have [P] [IA Option B 'suitable']'

R: Keep saying.

Igor: OK, [Line 5 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'have er [Option B] 'suitable' measurement methods' [P] hmm [IA 'and this one'] [P] [Line 6, in part, left-hand only, with KWIC inserted] 'efficient and biological biologically appropriate while [E] this solution [E] may be may not be [P] [IA Option B 'suitable']' [P] OK, I decide to choose er [Option B] B 'suitable'

R: Uh-huh.

Igor: I I think this er this question [E] all [E] [probably 'all of the options'] can apply in the [P] in the sentence but but the meaning is not appropriate.

R: Uh-huh.

Igor: So, I think you had to try all of them. The the way [the reason?] I I decide to choose [Option B] B 'suitable' because it looks make sense, yeah.

R: Er, why er again please? Why why do you choose B [Option B 'suitable'] rather than other choices?

Igor: When you translate, it's er it looks make sense.

R: Uh-huh.

Igor: Because another one is not make sense

R: Uh-huh.

Igor: But all of them you can apply it.

R: OK.

Igor: I don't know. Maybe may not [Option C] 'true', may not 'true'. But the meaning is looks strange when you put it.

R: Strange?

Igor: Yeah. The the grammar is correct but the meaning is [P] I I don't know what.

R: Are there any key words from Line 1 to Line 7 that you think that really [E] help [E] you?

Igor: Yeah, yeah, yeah. Some of them like [P]

R: For example?

Igor: [Line 6 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'may not be [Option B] 'suitable'? To [E] all [E] common property' something like that. It's looks psst [P] But anyway, [Option A] 'applicable' it's it can be the answer to this one.

R: No worries. Just, yeah, say it. Please say it.

Igor: [Line 4, in phrase, right before the KWIC position] 'socialist reform, a reform [KWIC position here]' [P] [Line 6, in phrase, right before the KWIC position] 'may not be [KWIC position here]' [P] Yeah, yeah, yeah, I like this one, Number 6 [i.e., Line 6]. I don't know.

R: Number 6? [chuckle]

Igor: Yeah [chuckle].

R: What what do you mean? It helps you with er [P] What are the key words in Line 6 that helps you?

Igor: [Line 6 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'may not be [Option B] 'suitable' [P] to all common property resource' Yeah, it's look [P] common. It's commonly used. I don't know I just guess, yeah.

R: Yeah, ok. Moment, you can tick that.

ConCloze 4: James on Item 2 (Target word: *applicable*)

Researcher: To your mind [P] whatever you are thinking

James: [Line 1] 'traits, which are characterized by the relative stability over time and [P] [KWIC position here] across situations, positive state-like capa capacities are relatively more like malleable and' [P] [Line 2, in part, left-hand only] 'urbanization display three research perspectives[:] abstract framework' [P] and [P] [Line 2 again immediately, in part] 'abstract frameworks that are universally blank [i.e., KWIC position] but at expense of the substantive speci [P] ficity[:] [P] context-dependent frameworks that' [P] [Line 3] 'require the use of resource that require [P] you clear blank [i.e., KWIC position] copyright [P] law before viewing and listening to them with your' [Line 4] 'aims for a breaking [in fact 'break'] just for socialist reform, a reform blank [i.e., KWIC position] in the new areas opened by the [P] revolution and its' [Line 5] 'still to be re resolved [P] and the [IA 'this er'] n374 [IA] something, eh? [IA] 134 have blank [i.e., KWIC position] measurement methods 43 have like methods 11 have potential or' [Line 6, in part] 'efficient and biologically appropriate. While this solution may not be [KWIC position here] to all common' [Line 6 again immediately, in part, from right before the KWIC position] 'may not be blank [i.e., KWIC position] all common property resources, for those of cases which may' [P] [Line 7] 'Court held that no long as a law in generally blank [i.e., KWIC position] and not discrimi [P] discriminatory in being specifically aimed at religion [P] [re]ligion' [P] the, OK, you want me to choose the [P] most suitable answer?

R: Yes. Now? [chuckle] Yeah.

James: [chuckle] er.

R: Keep saying whatever you are thinking.

James: Yeah, I'm, I have yeah three [P] potential answers, and just try to submit the those key words to these blanks and see er if it makes [IA] any sense. [Line 3, in part, left-hand only] 'require the use of resources require that [E] clear [E] that [in fact 'the'] [KWIC position here]' [P] so, [P] [IA] I think er [P] most of the cases the answer C [Option C 'true'] is not very suitable.

R: Uh-huh.

James: I mean er [P] I think it makes some sense but it's er in professional writing or something, it's not very good to use this [Option C] 'true'. So, either the [Option A] 'applicable' or [Option B] 'suitable' is going to be my choice.

R: Uh-huh.

James: Er [Line 4, in phrase, right before the KWIC position] 'just [IA 'for socialist'] reform, a reform [P] [KWIC position here] [IA]'

R: Keep saying.

James: Yeah, I just er I'm just er applying this [IA 'into sentences']

R: Yeah, yeah, just read whatever you are reading.

James: [Line 4 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'a reform [Option B] 'suitable' in the new areas opened by the revolution and its' [P] [Line 5, in part, with KWIC inserted] 'efficient and biologically appropriate. While the solution may not [E] be the [E] [P] [Option B] 's sh suitable' to all' I think its er [Option B] 'suitable' is OK, fine, I think [Option B] 'suitable' is the most er [P] suitable answer haha [laughing]

R: Haha [laughing], OK. A bit of a pun. OK.

James: Yeah, because er [P]

R: Because?

James: It's giving some literal sense, I mean, er apply this to this blank.

R: Which line?

James: All of them. All of them. The same.

R: All right. Yeah. Any key words in [E] particular [E] that help you?

James: No, in [E] generally [E] [P] er almost all to [IA] it's fine with [Option B] 'suitable' it's er if you say er [P] [Line 6, in phrase, right after the KWIC position, with KWIC inserted] '[Option A] 'applicable' measurement ['methods' not verbalized]', [Line 4, in phrase, right before the KWIC position, with KWIC inserted] 'a reform [Option A] 'appli[cable] [IA]', [Line 3, in phrase, right before the KWIC position] 'clear the [P] er [KWIC position here]', [P] [Line 2, in phrase, from right before the KWIC position, with KWIC inserted] '[E] universally [E] [P] [Option A] 'applicable' b [IA] but the but the' [P] I think it's er [P] if I er [Option A] 'applicable' and [Option B] 'suitable' are equally OK, but er according to my experience, the [Option B] 'suitable' is er most language some of this word [the most suitable one?]. That's why I use the. [P] Compared to this two, this two, giving good sense.

R: OK.

ConCloze 4: Klavier on Item 2 (Target word: *applicable*)

Researcher: Whatever you are thinking, whatever is going on in your mind. Just say it.

Klavier: Yeah, OK.

R: No worries about grammar or whatever.

Klavier: Right. [Line 1] 'traits which are characterized by relative stability over time and [P] [KWIC position here] across situations, positive state-like capacities are relatively more [P] malleable and' [P] [Line 2] 'urbanization display three research perspectives: abstract frameworks that are [E] universally [E] [KWIC position here] [P] but at the expense of substantive speci specificity; context-dependent frameworks that' [Line 3, in part] 'require the use of resources that require you [E] clear the [E] [KWIC position here] [P] copyright laws' I think [P] I think it's [Option A] 'applicable' [P]

R: Uh-huh.

Klavier: Because erm it can be like [Line 3 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option A] 'applicable' copyright laws' and [P] erm it can be [Line 1, in phrase, right after the KWIC position, with KWIC inserted] '[Option A] 'applicable' across situations'

R: Uh-huh. [P] So, you have read just three lines and then you [P] got the answer already?

Klavier: Erm, yeah, I think erm I think that kind of fits all. If it fits erm three answers [i.e., three concordance lines], then it should be correct. I mean, [IA 'three'] a bit [IA 'rest'] not quite enough.

R: So, why why do you think it er why do you think the answer is not [Option B] 'suitable' or [Option C] 'true'?

Klavier: Erm [P]

R: No worry, I would like to know.

Klavier: Yeah, [P] Erm, well, it kind of fits the contexts, I think. erm because it's like 'relative stability' [from Line 1] erm, and that kind of er 'a trait' [from Line 1] and the other [IA] [IA 'traits' [from Line 1] as well'], but [Option A] 'applicable' [IA 'cause'] 'stability' [from Line 1] and [P] being 'applicable' [i.e., Option A] is kind of similar? So, erm, [P] I think I think that's why.

R: What about the other lines? Are there any key words that help you [P] specially?

Klavier: Erm [P] for this one, [Line 4, in phrase, from right before the KWIC position, with KWIC inserted] 'a reform [Option A] 'applicable' in the new areas' erm [P] [Option B] 'suitable' wouldn't be correct. I don't think it's [Option C] 'true' as well. Just wouldn't be that's erm because you can apply them to to 'new areas' [from Line 4]. Erm [P] so, yeah [P]

R: OK. Yeah, good.

ConCloze 4: Lulu on Item 2 (Target word: *applicable*)

Lulu: Hmm, [P] [Line 1] 'traits which are characterized by relative stability over [E] time [E] and [KWIC position here] across situations, positive state-like capacities are relatively more [P] malleable and' [P] [Line 1 again immediately, in part] 'traits which are characterized by relative stability over time and the something [i.e., KWIC position] across situations' [P] psst, hmm, [P] [Line 1 again immediately, in part, left-hand only, with KWIC inserted] 'characters [in fact 'characterized'] by relative stability over time and the [P] [Option A] 'applicable', [Option B] 'suitable', [Option C] 'true' [P] hmm [P] interesting, because I was expecting [P] maybe a [E] noun [E] here in this blank. But somehow all these are [E] adjectives [E]. Er make me feel that erm [P] don't, they don't feel like they are parallel structures in these sentence. Hmm, OK. [P] er I'll just go for the next one. [Line 2, in part, with KWIC inserted] 'urbanization display three research perspectives: abstract frameworks that are [E] universally [E] [P] [Option A] 'applicable'? but at the expense of er [P] substantive specifility [in fact 'specificity']' [P]. From this sentence I feel like it should be [Option A] 'appli applicable'. It just er [P] [E] sound right [E]. haha [laughing] makes it's er it's er almost like you know. [P] If you talk about er [P] 'framework' [from Line 2], you talk you, er it's it's also it's almost like when you talk about framework, the erm connotation of 'framework' [from Line 2] [P] is almost, you know, you you [E] apply [E] [word of the same lemma as [Option A] 'applicable'] framework. [P] Right? So you say whether 'framework' [from Line 2] is [Option A] [E] 'applicable' [E]. You don't say whether er a a framework is [Option C] 'true'. Erm [P] you may say whether framework is [Option B] 'suitable'. But it's [Option B] 'suitable' to some kind of situation, it's almost like you need a [P] erm psst [P] you need a er erm [P] another erm [P] phrase [IA 'in the end'] to to for the qualify [Option B] 'suitable'. So, I feel like it should be [Option A] 'applicable'. So, I'll go [P] psst, to the next one. [Line 3, in part, with KWIC inserted] 'require the use of resources [P] that require you [E] clear [E] the [P] [Option A] 'applicable' copyright laws before viewing' hmm [P] er [Option A] 'applicable' here should be mean meaning, you know, relevant or er what law's applied here in this certain field. So, yeah, psst I think [P] I'll go for [Option A] 'applicable' for this one.

Researcher: OK.

Lulu: Hm, [P] next one.

Appendix 7. Verbal Reports for ConCloze 7 (Sections 4.3.2 and 4.3.3)

NB: Verbal reports to Item 2 (target word: *applicable*) are analyzed in both ConCloze 2–4 (Section 3.3.3 Testing usability and Section 3.3.4 Substantive content) and ConCloze 7 (Section 4.3.2 Test-taking processes and strategies). They have already been provided in Appendix 6 (pages 390ff.). To avoid redundancy, they will not be provided in this appendix. The verbal reports presented here are collected in ConCloze 2–4 and so entitled as such, even though used for substantive fine-tuning in Section 4.3.2 of ConCloze 7. They are grouped by the target words (*endeavor* in Item 1, *hypothesize* in Item 3, *recreational* in Item 4, and *livestock* in Item 5), and sorted by pseudonyms of the respondents.

ConCloze 2: Aaron on Item 1 (Target word: *endeavor*)

Aaron: [Item instructions] ‘You may need to maximize the web browser display to fit the question. Choose the most [E] appropriate [E] answer. [P] Which of the following should be that [E] word [E]?’ [P] Hmm.

Researcher: Keep talking.

Aaron: It’s a big, it’s a big question. And it has seven subquestions, so I need to, [P]

R: Just just say it.

Aaron: Yeah, I, s, s, er, s [Line 1, in part, left-hand only] ‘theoretical detail ultimately undermines the’ [P]

R: You don’t have to explain. Just [P]

Aaron: Yeah, yeah, yeah. I’ll have a quick [IA ‘guidance’] of the seven questions. And then I will look at the four choices, and I think I will need to put each of the choices back to the seven subquestions. [P] Hmm [P] [Line 1 continued, in part, from right before the KWIC position] ‘nature of the [KWIC position here]’ [P] [Line 2] [IA ‘their world but that’] ‘education in such a primary [E] human [E] [KWIC position here]’ [P] [‘as music should be universal for our’ not verbalized] [IA ‘children their’] [Line 3] ‘logical to go with your strengths and to [E] structure [E] [KWIC position here]’ [P] [Line 3 again immediately, in part, from right before the KWIC position] ‘[E] structure [E] [KWIC position here]’ [P] hmm [‘in such a way as to guarantee success’ not verbalized] ‘by anticipating’ [Line 4, in part, left-hand only] ‘and coordination across divisions of the college must occur. [P] hmmm [P] [KWIC position here]’ [sigh] [Line 5, in phrase, right before the KWIC position] ‘tedious and time-consuming [P] [KWIC position here]’ [IA] [P] OK, I think I could exclude er Choice D [Option D ‘fruitlessness(e)’] because er in the last two subquestions, they are plural forms. And the last number is er the last letter is s. But if I choose D [Option D ‘fruitlessness(e)’] [P] I think [P] Oh, this. [P] No, no, no, no. [P] They may also be appropriate because you add ‘e’ afterwards. Hmmm [P] I don’t know. The four choices are similar. Hmmm [P] My feeling is A [Option A ‘attempt’]. [P]

R: OK.

Aaron: Hahaha [laughing]. It takes a lot of time to answer only one question, so if I meet this kind of question, er and I have no er logic about how to answer this question, I will [P] just click one of the choices and leave it [P] because I will takes me a lot of time. If if in examination and it takes a lot of time for only one question, I will miss the other questions. And the other questions may be easier than this one. [P] So, I think I need to balance the difficulty of different questions.

R: Uh-huh. So, which words do you pay attention to specif er specially?

Aaron: Specially? Hmm [P] It’s it’s just er kind of feeling. When I read the sentence, I think er I will put each of the choice back to the sentence and er [P] and read in my mind to to think about whether it is fluently or not. If I think it’s it’s a little er [P] it’s [P] a little weird, I think it’s not appropriate. But it’s just the feeling. I didn’t have clear logic how to answer this kind of question.

R: So, you look at the choices and try to put [E] them [E]

Aaron: Yeah, back to the sub...

R: So, how can you know which choice is the correct one or the most appropriate one?

Aaron: Hmm, [P] I don't know. It's just er kind of feeling, yeah.

R: Alright, so the next question, please.

ConCloze 2: Björn on Item 1 (Target word: *endeavor*)

Björn: Question is t [Line 1] 'theoretical details ultimately undermines the properly interdisciplinary [P] nature of the erm [KWIC position here]' [P] [IA 'what'] I so put [P] in between them

Researcher: Yes.

Björn: Er [Line 1 continued] '[KWIC position here] and results in a dominance [P] of the'

R: So just say whatever, anything, that comes to your mind

Björn: [IA] hmm [chuckling] it's a bit it's a little bit difficult for me. Erm

R: Yeah, it's alright.

Björn: [Line 1 again immediately, in phrase, right before the KWIC position] '[IA 'interdisciplinary'] nature of the [KWIC position here]' [P] [Line 2, in word, right before the KWIC position] 'human [KWIC position here]'? Uh-huh? No.

R: Erm, these are the choices.

Björn: O, these are choices. [P] er [IA question stem 'all the lines above'] 'miss the same word' OK. Er So, I haven't, no, for

R: Yes,

Björn: Hmm, [Line 2 again immediately, in part, left-hand only] 'their world, but that education in such a primary human [KWIC position here]' or something similar [P] This's an [P] adjective? [IA] No, noun. [Line 2 again immediately, in phrase, right before the KWIC position] '[IA 'such a primary human'] [KWIC position here] [IA]'

R: Just say. Think and speak.

Björn: O, [Line 5, in part, right after the KWIC position] '[KWIC position here] that requires considerable [E] inconsequential [E] erm travel [P] over vast expanses of teri territory' and [Line 6] 'devoted to social activity would leave less time for [E] academic [E] and [KWIC position here] and therefore would likely have a negative impact on academic' [P] er [Line 7] 'being answered er but as is always the case in sci sci scientific [KWIC position here] and more questions have arisen. We haven't yet found evidence for' [P] erm [P] I think it's [IA]

R: Just say whatever that comes to your mind.

Björn: [Option D] 'fruitful fruitless fruitlessness, fruitlessness'? The answer is [Option D] 'fruitlessness'. Is what?

R: I dunno.

Björn: OK, so erm I just key in and pass it?

R: OK. Go on, go on.

ConCloze 2: Claire on Item 1 (Target word: *endeavor*)

Researcher: So, say it loud. Think aloud.

Claire: [Instructions] 'Choose the most appropriate answer.'

R: Whatever comes to your mind, just say it.

Claire: Hmm

R: What word you are reading, just say it.

Claire: I think I eh I'm reading er the sentence number one? So, [Line 1] 'theoretical'

R: Yes, good.

Claire: [Line 1 continued] 'detail ultimately undermines the properly [P] interdisciplinary nature of the [P] blank [i.e., KWIC position] and results in er dominance of the social sciences in' [Line 2, in phrase, left-hand only] 'their world'

R: Say it clearly and then go on.

Claire: Er

R: If you repeat any word, just say it again. Alright?

Claire: OK, and then er I think after reading the first two er [P] after reading the sentence on the number one [i.e., Line 1] [P] and then I went to see the question and then examples [P] or answers

R: OK, so

Claire: [Question stem] 'All the lines above miss the same word. Which of the following should be that word?' [P] Er I try to choose [P] er answer [P] hmm

R: Say it.

Claire: Yeah, [laughing] really difficult er to say and to solve the problem

R: OK [chuckle] try please.

Claire: [IA] [Line 1] 'theoretical detail ['ultimately' skipped] undermines [P] [some words not verbalized] nature of the [P] [KWIC position here] and results in the ['dominance' not verbalized] of the social sciences in' [Line 2] 'their world' anyway I will [P] er [IA 'turn'] to [P] the sentence in the number two [i.e., Line 2]

R: Uh-huh, good.

Claire: [Line 2 continued] 'but that education in such a [P] primary [E] human [E] [P] blank [i.e., KWIC position] as music [P] should be universal for our children [P] in [E] their [E]' [P] [Line 2 again immediately, in part, from right before the KWIC position] 'primary [E] human [E] [KWIC position here] and [P] as music should be universal [P] for our children [P] in [E] their [E]'

R: Good, good.

Claire: [Line 3, in word, left-hand only] 'logical' [P] I will continue er

R: Yeah, yeah, go on. Just say whatever you are reading.

Claire: [Line 3 again immediately] 'logical to go with er your strengths and to structure [P] [E] the [E] [P] the [KWIC position here] in such a way as to guarantee success by anticipating' [Line 4, in part, with KWIC inserted] '[IA 'and coordination across divisions of the college'] must occur. [P] This [IA [Option A] 'attempt'] is a [IA 'campuswide effort']' [Line 5] [IA]

R: Say it, loudly.

Claire: [Line 5, in phrase, from right before the KWIC position] 'time-consuming [P] [KWIC position here]' hmm [IA]

R: Go on. Say it, whatever you are reading.

Claire: This is er difficult reading

R: Yeah, just say it.

Claire: Yeah, [Line 5 continued] '[KWIC position here] that requires [IA 'vast expenses] of territory' [Line 6] 'devoted to social activity would leave less time for academic [P] [KWIC position here] er and therefore would likely have a negative impact' [IA] [Line 7, in part, from right before the KWIC position] '[IA] case in scientific [P] [KWIC position here] questions have arisen' [P] I think the answer is [P] [Option D] 'attempt'? [P]

R: Why?

Claire: Er [P] hmm [P]

R: Say it. [P] Why?

Claire: Hmm.

R: Say it. Think aloud. Do not think quiet. [chuckle]

Claire: Er [Option A] 'attempt.' [P] [Line 2, in phrase, right before the KWIC position] 'such ['a' not verbalized] pri[mary] [E] human [E] [P] [KWIC position here]' [Line 2 again immediately, in word, right before the KWIC position] 'human [P] [KWIC position here]' hmm

R: You said 'human'? Say it. Say it whatever you are thinking.

Claire: Erm, because I think er the word the answer requires it's more er [IA 'palative'??] behavior?

R: What? What behavior? Sorry?

Claire: [IA 'Palative'??] er attitude er [P] word [IA] [P]

R: Say it.

Claire: [IA] [Line 1, with KWIC inserted] 'theoretical detail [IA] interdisciplinary nature of the [P] [Option A] 'attempt' in such a dominance [P] social science er [P] in' [Line 2, in part, with KWIC inserted] 'their world but that education in such a primary human [P] [KWIC position here] er [P] [IA]'

R: Say it loudly.

Claire: [IA]

R: Think aloud, please.

Claire: I think it's a the answer is er one [Option A] 'attempt'

R: Uh-huh. Why?

Claire: Because er the first sentence er the [Line 1] 'theoretical detail undermines the properly interdisciplinary' [P] hmm [P] 'nature of the blank [i.e., KWIC position] and results in the dominance the social sciences in' [Line 2, in phrase] 'their world' [P] so, [P] er I think er it's er social science. It try to hmm [P] try to investigate [P] something? Across the field? [P]

R: Uh-huh.

Claire: So, I think [P] er the similar word with er contextual sentence?

R: What about the other sentences? Which words help you to get to the answer?

Claire: Er, psst. Er.

R: Any word that you think helps you specifically?

Claire: Hmm [Line 2 again immediately, in part, from right before the KWIC position] 'primary human [P] [KWIC position here] as music should be [P] universal for our children in their' [Line 3] 'logical to go with ['your' not verbalized] strengths to structure the [P] [KWIC position here] in such a way [P]'

R: Say it.

Claire: [Line 3 continued] 'by anticipating' [Line 4] 'and coordination across divisions'

R: These lines are from different places. Uh-huh. They are from different places. But they miss the same word.

Claire: [Line 4 continued] 'must occur [P] hmm [KWIC position here]' I think er [P] in the last two sentences, the last two parts of the first sentence [Line 3, in part, right-hand only] '[KWIC position here] in such a way as to guarantee success by er [P] participating [in fact 'anticipating']' [Line 4, in part, left-hand only] 'and coordination across divisions of the college must occur [P] [KWIC position here]'

R: These lines are not connected. Alright? They are from different places. They are not from the same passage.

Claire: [P] Hmm [P] But, you know, it's the, the first sentence [i.e., Line 1] it ends on the number five [i.e., Line 5]

R: No, they are not connected; these sentences are from different places. One to seven [i.e., Lines 1–7] are from different places.

Claire: [P]

R: So, anyway, why do you choose [Option A] 'attempt'? Which words in each line help you?

Claire: Hmm, [P] I think it's er it's the number one. [P] every er [P] every every the words it's er, it's er there is er there is no relation

R: No relationship to each other but they miss only they miss the same word. They are from different places.

Claire: Ah, so this is not a passage.

R: Not a passage. They are from different passages.

Claire: OK. [P] so, ev, so every is the word every number? Is there any [P]

R: No interconnection. The only connection is that they miss the same word.

Claire: Yeah, I see. This is why I felt very er difficult.

R: Yeah.

Claire: Hmm, [Line 1, in part, from right before the KWIC position] 'nature of the [KWIC position here] results in a dominance social [E] sciences [E]' [P] OK, I think in the in the number one [i.e., Line 1], I think it's er the clue is 'dominance of the social sciences' and 'interdisciplinary' [both are phrases from Line 1]? And 'undermine' [from Line 1]?

R: Uh-huh.

Claire: Yeah,

R: What about other lines? Are there any context clues to help you?

Claire: [Line 2, in part] 'but that education in such a primary human [KWIC position here] as music' [P] psst 'should be universal for our children in their' [P] hmm [P] [Line 2 again immediately, in word, right before the KWIC position] '[E] human [E] [KWIC position here]' [P] [Line 2 again immediately, in word, right before the KWIC position] '[E] primary [E] [KWIC position here]' [P] hmm

R: Say it. Whatever you are thinking. Just say it. No worries about grammar

Claire: I think it's er difficult to find any clue in the number two [i.e., Line 2].

R: Yeah, what about other lines? If you do not find, just go on.

Claire: Yeah. Number three [Line 3] 'logical [P] to go with [IA] strengths and er structure the [KWIC position here] in such a way as to guarantee success by anticipating' hmm

R: Say it, whatever you are thinking.

Claire: Hmm, [Line 3 again immediately, in word, right before the KWIC position] 'structure [KWIC position here]' [Line 3 again immediately, in part, from right before the KWIC position] 'structure the [P] [KWIC position here] anticipating' [P] erm I'm not sure. I think I need to pass, [P] to the number four.

R: Yeah. But just say it, whatever you are reading, thinking

Claire: [Line 4, in part, from right before the KWIC position] ‘coordination across divisions [IA ‘of the’] college must to [P] occur. This [P] [KWIC position here] [IA] campuswide effort’ [P] er I think [P] I think er [P] I think number four ‘occur’ [from Line 4]? Is er, is a word for hint. So, [Option A] ‘attempt’? Yeah. ‘Effort’ [from Line 4]?

R: OK, any other con

Claire: ‘effort’ and ‘occur’ [from Line 4] [P] number five is the [Line 5, in part] ‘adventure travel. It is a tedious and [E] time-consuming [E] er [KWIC position here] [P] that requires [IA]’ [P] I think erm [Line 5 again immediately, in phrase, right before the KWIC position, with KWIC inserted, re-structured] [Option A] ‘attempt’ its ‘attempt’ is er ‘tedious and time-consuming’. Yeah.

R: Say it. Whatever you are thinking.

Claire: They could be er [Option D] ‘fruit[lessness]’ for [P] [Option A] ‘attempt’ [P] and then next [Line 6, in part, left-hand only] ‘devoted to social activity would leave less time for academic [P] hmm [KWIC position here]’ [P] yeah [Line 6, in part, from right before the KWIC position, with KWIC inserted] ‘academic [Option A] ‘attempt’ [P] and therefore would likely have a negative impact on academic’ hmm

R: Any other contextual clues?

Claire: [Line 7, in part] ‘being answered but as is always the case in [E] scientific [E] blank [i.e., KWIC position] more questions have arisen. [IA]’ ‘found’? Yeah, so the hint is ‘found’ [from Line 7].

R: Why? I would like to know what you are thinking. No worries.

Claire: ‘Find’ the [P] ‘find’ [the same lemma as the word in Line 7] er could be similar word [Option A] ‘attempt’?

R: Uh-huh. Any other thing you would like to say? What helps you to get to the answer? What you are thinking? Any hypothesis you are forming?

Claire: Psst. Yeah, I think yeah.

R: OK.

ConCloze 2: Dakota on Item 1 (Target word: *endeavor*)

Researcher: You are going to choose the most appropriate answer for these seven lines. [The concordance lines are being indicated] [Question stem] ‘All the lines above miss the same word.’ So, ‘which of the following should be that word?’ [The options are being indicated.] OK? Just say whatever you are thinking, how you get to the answer.

Dakota: [Item instructions] ‘Choose the most appropriate answer.’ [P] So, for for this is just er lines, right?

R: Yes,

Dakota: But do they have they have seven

R: They miss the same word.

Dakota: O, it’s the same word.

R: The same word in er from different places.

Dakota: OK.

R: These seven lines are from different places.

Dakota: [Line 1, in part] ‘s theoretical detail’

R: You can use the mouse to scroll back [i.e., scroll left and right] yeah.

Dakota: [P] [P]

R: Say it please. Think aloud.

Dakota: But I can't. When I read this, I can't think.

R: It's OK that you read out loud so that I know what word you are reading.

Dakota: O [Line 1 continued] 'nature of [E] the [KWIC position here] and results [E]' [P] [E] the [E], after 'the' should be the should be er [E] noun [E], I was thinking [P] noun [Line 1 again immediately, in part, from right after the KWIC position] '[KWIC position here] and results'. Should be like er same to the er 'results'

R: Uh-huh.

Dakota: [Line 1 continued] 'in a dominance of the social' Hmm?

R: You can you can scroll here [indicating the scroll bar].

Dakota: Scroll here. Oh, OK. [P] [Line 1 again immediately, in phrase, right-hand only] 'of the social sciences in' [Line 2] 'their their world but that education in such a a primary human [KWIC position here]' [P]

R: Say it.

Dakota: [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'human [P] [Option A] 'attempt' [Option B] 'difficulties' [P]' [Line 1, in phrase, right after the KWIC position] '[KWIC position here] and results' [P] hmm [Line 2, in phrase, right after the KWIC position] '[KWIC position here] as ma as music' [P] [Line 2 again immediately, in part, from right before the KWIC position] 'human what? [i.e., KWIC position] as er [E] music [E] should be [P] universal for our children in their' [Line 3, in part, left-hand only] 'logic to go with your strengths and to structure [E] the [E] [KWIC position here]' [P] [Line 3 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'structure the [Option B] 'difficulty'' [P] What is [Option D] 'fruitlessness'? [P] [Line 1, in phrase, right before the KWIC position, with KWIC inserted] 'nature of the [Option D] 'fruitlessness' and' [P] [IA]

R: Say it. Say it, whatever you are thinking.

Dakota: [Line 4, in phrase, from right before the KWIC position] 'The er [E] This [E] [KWIC position here] [P] is er a campuswide effort ['that' not verbalized] [P] begun with er [IA] [P]' [Option A] 'attempt' [Line 4 again immediately, right before the KWIC position, with KWIC inserted] 'this [Option A] 'attempt' is' [P] [Line 5, in word, right before the KWIC position] 'time-consuming [KWIC position here]' [P] s [Line 5 again immediately, in phrase, right before the KWIC position] 'and er time [KWIC position here]' [P] These [P] seven should be one word. Same

R: Yes, but these seven lines are [E] not [E] from the same text. They are from different texts.

Dakota: I see what you mean. [Line 5 again immediately, in part, left-hand only] 'It is a [P] [IA] '[time-consuming [KWIC position here]'

R: OK, just say it, yeah. No worries whether your answer is right or wrong. Just say.

Dakota: [Line 5, in word, right after the KWIC position] '[KWIC position here] that' You want me to give answer now?

R: Up to you. You can read all lines. Or you can

Dakota: [Line 6, in part, left-hand only] 'would leave less time for [E] academic [E] [KWIC position]' [P]

R: Say it.

Dakota: And [P] What's what's this word [IA [Line 5] 'vast expanses'] [Option C] 'e-ven endeavor'?

R: You don't, it's ok, so you do not know the word in Choice C [Option C 'endeavor']? Uh-huh.

Dakota: [IA] [P] [Line 7, in part, left-hand only, with KWIC inserted] 'in ['always' not verbalized] [P] is the case in scientific [P] [Option B] 'difficulty' [Line 4, in word, right before the KWIC position, with KWIC inserted] '[E] this [E] [Option B] 'difficult[y]'' [Line 4 again immediately, in word, right before

the KWIC position, with KWIC inserted] 'this [Option B] 'difficult[y]'' [P] [Line 4 again immediately, in word, right before the KWIC position, with KWIC inserted] '[E] this [E] [Option B] 'difficult[y]'' [P] [Line 2, in word, right before the KWIC position, with KWIC inserted] 'human [Option B] diffi difficult[y]' [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] 'the [Option B] 'difficult[y]' and results' [Line 5, in phrase, from right before the KWIC position, with KWIC inserted] 'time-consuming [Option B] 'difficult[y]' that requires considera[ble]' [IA] [IA '[Option A] 'attempt'''] [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] 'time-consuming [Option A] 'attempt' [IA]'

R: Say it, whatever you are thinking. The way you think.

Dakota: [Line 6, in word, right before the KWIC position, with KWIC inserted] 'academic psst [P] academic [Option A] 'attempts'' [Line 6 again immediately, in part, from right before the KWIC position] 'academic [P] [KWIC [position here] and therefore would likely have a negative impact on academic' [P] [Line 7, with KWIC inserted] 'being answered but as is always the case in scientific [P] [IA] [Option A] 'attempts' [P] more questions have ['arisen' not verbalized] [P] We haven't yet found evidence for' [P] Ahh, I don't know. I just I just kind of choose one then.

R: Yeah, it's ok.

Dakota: I don't know. [P] no, no, sorry, this one maybe. Psst. O, dunno. I just confused.

R: Yeah, I would like to know how, yeah, the way you can get to the answer. So, why do you think it should be [Option B] 'difficulty'?

Dakota: This over, this all [IA 'noun'], right? [Line 1, in phrase, right before the KWIC position, with KWIC inserted] 'nature of the [Option B] 'difficulty' s' [P] I think it's just er [Option B] 'difficulty' it's just more suitable for [P] for

R: For all the lines?

Dakota: For all the lines. Sometimes it's put the it's put the C [Option C 'endeavor'], doesn't work with the [Line 2, in phrase, right before the KWIC position] '[IA 'primary'] human er [KWIC position here]' [P] that's for example, A [Option A 'attempt'] [P] [Line 2 again immediately, in word, right before the KWIC position, with KWIC inserted] 'human [Option A] 'attempt'' doesn't sound right. It's doesn't sound right, you know.

R: Uh-huh. [P] What's the

Dakota: [Option B] 'difficulty'

R: What do you mean 'Doesn't sound right'?

Dakota: Doesn't sound right? [P] Hey? That's, I mean, [P] doesn't er psst [P] doesn't er this word [P] [Option A] 'attempt' doesn't go with the [P]

R: Go on.

Dakota: Doesn't go with the some, like er, some words [indicating on the screen]

R: 'academy', you mean?

Dakota: Yeah yeah 'academic' [from Line 6].

R: You mean 'academic'[from Line 6] and [Option A] 'attempt'? Do not go together?

Dakota: Yeah, yeah. [Line 6, in word, right before the KWIC position, with KWIC inserted] 'academic [Option A] 'attempts''? er the [P] [Line 5, in word, right before the KWIC position, with KWIC inserted] 'time-consuming [Option A] 'attempt''? I I er [P] I dunno. [P]

R: Alright.

Dakota: I think it's the [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] 'nature of the [Option A] 'attempts' and results'? I dunno. Just doesn't doesn't sound right. I just feel I just feel the words.

R: Alright,

Dakota: I mean, the following words [the options] doesn't sound right, you know.

R: So, er, the word before and then the blank, you you think they

Dakota: After

R: You think they [P] You try to match them, right?

Dakota: Yeah, yeah, yeah.

R: Ow, alright. [P] Anything else that you would like to say?

Dakota: No.

R: OK.

ConCloze 2: Esther on Item 1 (Target word: *endeavor*)

Esther: Right. OK, I have to choose the most appropriate answer.

Researcher: Yes, you can use the mouse so that you can scroll.

Esther: And I have to read this statement?

R: Yeah. [P] These lines are from [P] different places.

Esther: Right.

R: You need to read this and then you choose an answer.

Esther: OK, [Line 1] 'theoretical detail ultimately undermines the properly [P] interdisciplinary nature of [KWIC position here]' [P] Oh, ok, so it's one for sentence[s] that I have to fill in the blanks over here.

R: Yes, that's right. But these lines miss the same word.

Esther: I have to choose one word from here.

R: Yes, that's right. You have seven lines, and these seven lines miss only one word. They miss the same word.

Esther: Alright. OK.

R: Yeah.

Esther: [Line 1, in part, from right before the KWIC position] 'nature of the dash [i.e., KWIC position] and results in the dominance of social sciences'. Right.

R: Yeah.

Esther: [Line 2] 'their world, but that education in such a primary [E] human [E] [KWIC position here]' [P] Right. [Line 2 continued] '[KWIC position here] as music should be universal for our children' [Line 3] 'logical to go with your strengths and to structure the dash [i.e., KWIC position] in such a way as to guarantee success by anticipation? [P] Anticipating?' Can I make it a bit smaller, as I could not see it the whole thing in one screen?

R: Uh-huh, sure. Let me see [P] [Screen adjusted]

Esther: So, the four words are [Option A] 'attempt', [Option B] 'difficulty', [Option C] 'endeavor', and [Option D] 'fruitlessness', right. Erm [Line 4] 'and coordination across divisions of the college must occur. [P] The dash [i.e., KWIC position] is a campuswide effort [P] that begins with a coordinating committee' [P] Uh-huh. [P] Right. Over here [referring to Line 4] [Option C] 'endeavor' fits well. [P]

R: Uh-huh.

Esther: Erm, [Line 5, in part, with KWIC inserted] 'or adventure travel. It is a tedious and time-consuming [P] [Option C] 'endeavor' that requires considerable inconsequential travel across [in fact 'over'] [P]

Fine. [Line 6, with KWIC inserted] ‘devoted to social activity would leave less time for [P] academic [Option C] ‘endeavors’ and therefore would likely [P] have a negative impact on academic’ [P] These are not complete sentences. They are just part

R: Yeah, yeah, yeah. That’s right. It’s a special kind of language test.

Esther: Alright, ok. [Line 7, in part, with KWIC inserted] ‘being answered but as is always [P] the case in scientific [Option C] ‘endeavors’, more questions have arisen.’ I think [Option C] ‘endeavor’ a bit more.

R: So, why why do you think it should be [Option C] ‘endeavor’? What are the clues that you get?

Esther: Just because it makes sense [IA ‘whatever’]. You know, these sentences are not complete. They are part of sentences [IA]. So, just [IA] Like in the first one, I’m thinking like er I’m not very sure which will feel fit [P] but this sentence [i.e., Line 4]

R: Line 4?

Esther: Because this is full-stop here.

R: Uh-huh.

Esther: And then this sentence [still Line 4]. Because there is the subject and then [IA] [P] the sentence structure [i.e., ‘This ___ is a campuswide effort...’] is such that from here I pick up that that it could be [Option C] ‘endeavor’. [Line 4 again immediately, from right before the KWIC position, with KWIC inserted] ‘This [Option C] ‘endeavor’ is a campuswide effort that begins’

R: Uh-huh.

Esther: And you see over here: also it makes sense, because it’s the beginning of the sentence. I think the first three sentences [i.e., Lines 1–3], I don’t have the beginning, so I don’t what is the subject. That’s why it’s difficult for me to make out what erm like what’s the theme of the sentences. But in [E] this [E], next one

R: Four and Five [Lines 4 and 5]?

Esther: There is a full stop here [Line 4]. There’s the [Line 5, in phrase, left-hand only] ‘or adventure travel’ does not make sense, but there is the full stop, and then it says [Line 5 again immediately, in part, left-hand only, with KWIC inserted] ‘it is a tedious and [E] time-consuming [E] [P] [Option C] ‘endeavor’’, so like I can say

R: Why why do you think it should be [Option C] ‘endeavor’, not [Option B] ‘difficulty’, [Option D] ‘fruitlessness’ and [Option A] ‘attempt’ for Four and Five [i.e., Lines 4 and 5]?

Esther: Well, because you said you said, they have to come everywhere. So, this [P] Yeah OK, it could be [Option A] ‘attempt’ also, [Line 4, in part, from right before the KWIC position, with KWIC inserted] ‘This [Option A] ‘attempt’ could be a nation er campuswide effort’ actually

R: Because I wish to learn how you get to the answer.

Esther: Er [Line 5, in part, with KWIC inserted] ‘It’s a tedious and time-consuming [Option A] ‘attempt’’ Yeah, that could also be. [P] [Line 5 continued] ‘[KWIC position here] that’s requires considerable.’ Yeah that’s right. Word ‘attempt’ could be there. Er in this one [P] [Line 6] ‘devoted to social activity would leave less time for academic [P] [Option A] ‘attempts’ and therefore would likely have a negative impact on academic’ [P] You know the problem is because the sentence has [P] I don’t know the beginning. I don’t know the end. [P] It’s in the middle. So, it’s more of a guesswork.

R: But, yeah. But still you you know there is [P] in it. You see the possibility

Esther: Yes, [Option B] ‘difficulty’, I think grammatically it won’t fit very well because er especially [P]

R: You mean, in Line 4? In Line 4, the word ‘difficulty’ [Option B] and ‘fruitlessness’ [Option D]

Esther: Yeah, because you see, if you say [Line 4, in part, from right before the KWIC position, with KWIC inserted] ‘This [Option D] ‘fruitlessness’ is a campuswide effort’. Because [Option D] ‘fruitlessness’ is not an ‘effort’. That’s why it will not make sense here.

R: Uh-huh, uh-huh.

Esther: But you can say [Line 4 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'This [Option C] 'endeavor' is' because that could be [P] 'the effort' [from Line 4], and this [Option A] 'attempt' can also be 'the effort' [from Line 4] so that's why.

R: Uh-huh, interesting. Anything else that you would like to say about this question?

Esther: It's difficult to understand sentences when you don't have the beginning haha [laughing]

R: But still you can figure out the answer

Esther: You know, I just pick it up from here, because you see, this one, [P] and also in this one.

R: Line 4 and 5 [sic].

Esther: It's the beginning. When you have the subject kind of thing, and then

R: It's easier.

Esther: It's making more sense.

R: It's easier to get the answer when you get the sentence beginning somewhere.

Esther: Yes.

R: O, right. This is interesting. Thank you.

ConCloze 3: Franz on Item 1 (Target word: *endeavor*)

Researcher: OK, say whatever comes to your mind.

Franz: Erm, [P] things [P] that are coming to my mind is how to translate this, again, erm, [Line 1, in part, left-hand only] 'theoretical detail ultimately undermines the properly interdisciplinary nature of [E] the [E] [KWIC position here]' So it must be noun, [P] right? Coz I can see the article [i.e., the definite article 'the']

R: Uh-huh.

Franz: Right? [Line 1 again immediately, in part, from right before the KIWC position] 'the nature of the something [i.e., KWIC position] and results erm in the dominance of the social sciences in tut tut tut tut' [P] [Line 2, in part, from right before the KWIC position] 'in such a primary [E] human [E] tut tut tut [i.e., KWIC position] [P] in their' [Line 3, in part, left-hand only] 'logical to go with your strengths and to structure the [KWIC position here]' [IA] hmm its must its must be noun [Line 3 continued] '[KWIC position] in such a way as to guarantee success' [P] erm [P] OK, so erm. Can I choose now?

R: Up to you.

Franz: Hmm [P] let's translate some more sentence hahaha [chuckle]

R: OK [chuckle]

Franz: [NB: Line 4 appears to be skipped altogether or at least not verbalized here.] [Line 5, in part] 'or adventure travel. It is a tedious and [E] time-consuming [E] tut tut tut [i.e., KWIC position] [P] erm [P] that requires [IA 'considerable'] inconsequential'

R: Keep saying.

Franz: Erm [Line 6] 'devoted to social activity would leave less time for academic tut tut tut [i.e., KWIC position] and therefore would be likely to have negative impact [real text somewhat different] on academic' [Line 7, in part] 'being answered, but as is always the case in scientific hmm [KWIC position here] more questions have arisen.' Psst, [P] well, well, well, [P] so, erm, [P] for now, I try to guess from the clue as you, er as we did before. Erm [P] it must be something that can be [E] structured [E] because it see the word 'structure' [from Line 3]

R: Yes,

Franz: And [P] the [E] verb [E], [P] oh the noun itself must be something erm [P] er that [P] that is a kind of process? Because it can be 'time-consuming' [from Line 5].

R: Uh-huh.

Franz: Right?

R: I dunno.

Franz: Hahaha.

R: Just keep saying, yeah?

Franz: Well, [P] hmm [P]

R: Whatever you are thinking, whatever you are reading

Franz: Er, I am thinking about the er meanings erm erm the words you provided here erm the answer A [Option A 'attempt'] and the answer C [Option C 'endeavor'], to some certain degree, it can be synonymous. Am I right? erm [P] And [P] [IA 'See it? This is my competency!'] hahaha [laughing] And Choice B [Option B 'difficulty'] and Choice D [Option D 'fruitlessness'] erm [P] can potentially be synonymous as well.

R: OK.

Franz: So, [P] you you can choose among these two pairs.

R: Uh-huh.

Franz: Right? Erm [P] hmm [P]

R: Keep saying whatever you are thinking

Franz: [P] erm [P] what I'm thinking is erm it must be [P] A [Option A 'attempt'] or C [Option C 'endeavor']

R: Uh-huh

Franz: But erm psst [P] but C [Option C 'endeavor'] sound more formal [as opposed to Option A 'attempt']

R: Uh-huh

Franz: Compared to A [Option A 'attempt'], I guess. [P] So, in this context erm [P] the the whole passage or message or whatever [chuckle] from number one to number seven [i.e., Lines 1–7] I can see that it's talking about erm [P] 'academic' [word from Line 6], 'scientific' [word from Line 7], 'structure' [word from Line 3], 'human' [word from Line 2], 'nature' [word from Line 1] erm [P] which [P] could be [E] more formal [E]

R: Uh-huh.

Franz: Erm, [P] so choose, I do choose the answer C [Option C 'endeavor'], OK?

R: OK. Yeah, you can click it. [Franz clicking the answer] And anything else you would like to say about that one? Nothing more?

ConCloze 3: Gill on Item 1 (Target word: *endeavor*)

R: OK, just say whatever you are thinking.

Gill: OK, erm er [Line 1] 'theoretical detail ['ultimately' not verbalized] undermines the properly interdisciplinary nature of [E] the [E] [KWIC position here] [P] and results in a dominance of the social sciences' [P] er [IA] [IA Option D 'fruitlessness'] [P]

R: Just keep saying.

Gill: [Line 2, in part, left-hand only, with KWIC inserted] 'their their world but that education in such a primary ['human' not verbalized] [P] [Option C] 'endeavor' [P]'

R: Keep saying.

Gill: [Option C] 'endeavor', 'human' [from Line 2], [Line 2 again immediately, in word, right before the KWIC position, with KWIC inserted] 'human [Option C] 'endeavor'', [Line 1, in phrase, right before the KWIC position, with KWIC inserted] 'nature of the [Option C] 'endeavor'' [Line 3, in word, right before the KWIC position] 'structure [KWIC position here]', [Line 3 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'and to structure [Option C] 'endeavor'' [P] er [P] [Line 5, in part, left-hand only, with KWIC inserted] 'or adventure travel. It is a tedious ['and' not verbalized] time-consuming [P] [Option C] 'endeavor'?' [P] er [Line 6, in part, left-hand only, with KWIC inserted] 'devoted to social activity [P] would leave less time for academic [P] [Option C] 'endeavor [P] endeavors' [Line 7, in part, left-hand only, with KWIC inserted] 'being answered but is always the as is as is always the case in scientific [Option C] 'endeavors''. I think [Option C] 'endeavor'. It's C, [Option C] 'endeavors'. Yeah.

R: Uh-huh, so

Gill: [Line 1, in part, left-hand only] 'theoretical detail ultimately undermines the properly interdisciplinary nature of the [Option C] 'endeavor' [P]'

R: So, you think the answer is [Option C] 'endeavor'?

Gill: Yeah.

R: OK, you can tick it, and

Gill: [Clicking on the box of Option C 'endeavor']

R: Could you tell me again, moment, could you tell me again what are the key words that help you? The context clues that help you to get to this answer?

Gill: Er, ok, yeah.

R: What did you use and [E] how [E] did you do it?

Gill: Er, er [P] [Line 1 again immediately, in part, left-hand only] 'ultimately undermines [P] the proper[ly] inter' here the first one [IA 'was'] er 'interdisciplinary' [from Line 1].

R: 'Interdisciplinary' [from Line 1]? Uh-huh.

Gill: Yeah, it helps me to get the word. [P] Yeah, 'interdisciplinary' [from Line 1]. And then it's, sorry, that that is the 'nature', 'nature' [from Line 1].

R: Ah, 'nature' [from Line 1].

Gill: 'Nature', yeah, 'nature' [from Line 1]

R: 'Nature of [Option C] 'endeavor'?'

Gill: [Line 1 again immediately, in phrase, right before the KWIC position] 'nature of [KWIC position here]' yeah.

R: What about other lines?

Gill: Then the second one was er [Line 2, in part, left-hand only] 'their world, but that education in such a primary [E] human [E] [KWIC position here]

R: Ah.

Gill: [Line 2 again immediately, in phrase, right before the KWIC position] '[E] primary human [KWIC position here] [E]' also helps me to get. Then the third one, [Line 3, in part, left-hand only] 'logical to go with your strengths and to str str structure [KWIC position here]' [Line 3 again immediately, in phrase, right before the KWIC position] '[E] 'structure the [KWIC position here]' [E] [Line 3 again immediately, in phrase, right before the KWIC position] 'structure the [KWIC position here]' [IA 'to way'] 'structure' [from Line 3] that help me. And then er the the [Line 4, in part, left-hand only] 'coordination across the divisions of the college must occur. [E] This [E] [KWIC position here]' yeah the the direct the direct er the definite article 'this' yeah 'this' [from Line 4] helps me.

R: What what about, why why don't you think it could be [Option A] 'attempt', [Option B] 'difficulty', or [Option D] 'fruitlessness'?

Gill: Er

R: What tells, what tells you, yeah?

Gill: Hmm [P] erm [P] this [P] the [Line 4, in phrase, from right before the KWIC position] 'This [KWIC position here] is the campuswide effort' [P] yeah, because [P] if you if you go to the next line,

R: Uh-huh, the fourth one [i.e., Line 4].

Gill: Yeah, yeah, yeah the fourth one. It's the 'the effort' [from Line 4], you know, 'the effort', it's it must be doing something that is to produce some effort. There must be a linkage.

R: Uh-huh, uh-huh, good.

Gill: OK, so, it cannot be [Option B] 'difficulty'

R: Ah.

Gill: Yeah, it cannot also be [Option A] 'attempt'. And it can't be [Option D] 'fruitlessness'.

R: Ah.

Gill: So, the appropriate word is the [Option C] 'endeavor'. Also the s the subsequent use of the word 'effort' [from Line 4] then you go to [IA 'on'] er [Line 5, in part, left-hand only] 'or adventure travel. It is a tedious and time-consuming [KWIC position here]' [P] You can see, [Line 5 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'it's time-consuming [Option A] 'attempt''

R: Uh-huh.

Gill: [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] 'time-consuming [Option B] 'difficulty', or [Line 5 again immediately, in phrase, right before the KWIC position] 'it's time-consuming [KWIC position here]' [P] It will be [Line 5 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'it's time-consuming [Option C] 'endeavor''

R: Ah. [P] So, let's say, from the first five lines, you can get to the answer.

Gill: Yeah, you can, you can.

R: Ah.

Gill: Yeah, sometimes [IA 'add up'] from the [IA 'who's this?], or from a particular group of words or the the result of action.

R: Ah.

Gill: OK, now, the sixth one, [Line 6, in part, left-hand only] 'devoted to social activity would leave less time for [E] academic [E] [KWIC position here]' yeah, if you are leaving less time for some day, it means that it might be something that is doable.

R: Ah.

Gill: And so, [Line 6 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'academic [Option C] 'endeavor endeavors' and therefore would likely have a negative impact on academic', OK, then the seventh one [Line 7, in part, left-hand only] 'being being answered but as is always the case' [P] [Line 7 again immediately, in part, left-hand only] '[E] as [E] [P] is always the case in scientific sci [KWIC position here]' of course it cannot be [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] 'scientific [Option A] attempt', or [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] 'scientific [Option B] 'difficulty' or [Option D] 'fruitlessness''. It must be it's [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] 'scientific [Option C] 'endeavor''.

R: Wow, OK. Anything else that you would like to say? Or,

Gill: Yeah, erm, [P] [IA 'you know that the'] the sometimes you there is the you are likely to choose [Option A] 'attempt'

R: Yeah.

Gill: Because it's it's it's it's somehow [IA 'being'] leads a bit to [Option C] 'endeavor'.

R: O.

Gill: So, if you are not careful, yeah, it's it's clues words to [Option C] 'endeavor'.

R: Yeah.

Gill: Yeah, yeah, but I I think that it it is there to confuse. They

R: Yeah, and that is the correct answer. Perfect.

ConCloze 3: Halle on Item 1 (Target word: *endeavor*)

Halle: OK, so, I have to choose the most appropriate answer for seven different sentences. So, I have to choose the same word for all of them. So the first sentence is [Line 1] 'theoretical detail ultimately undermines the properly interdisciplinary nature of the [E] blank [E] [i.e., KWIC position] and results in a dominance of the social sciences in' [Line 1 again immediately, in phrase, right before the KWIC position] 'nature of [E] the [E] [KWIC position here]' well, [P] for this one I think [P] the word that would fit [E] might be [E] [P] [Option C] 'endeavor' [P] or [Option A] 'attempt'. So I'm gonna go

R: Yeah.

Halle: To see the next one coz it's too too soon to say.

R: Yeah.

Halle: So, the next one would be [Line 2] 'their world, but that education in such a primary [E] human [E] blank [i.e., KWIC position] as music should be universal for our children in their' [P] again, for this second one, I would say [Line 2 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'pri such a primary human [Option C] 'endeavor'' because I don't think [Option B] 'difficulty' makes sense, or [Option D] 'fruitlessness' [P] or [Option A] 'attempt'. So, for now, I'll go with [Option C] 'endeavor', but [IA 'will'] see if it fits [P] er go for the other ones.

R: Yeah.

Halle: So, the third one is [Line 3] 'logical to go with your strengths and to structure the blank [i.e., KWIC position] in such a way as to guarantee success by anticipating' [P] Again I think [Option C] 'endeavor' fits this sentence, so I'm gonna go to the next one. [Line 4] 'and coordination across divisions of the college must [E] occur. [E] This blank [i.e., KWIC position] is a campuswide effort that begins with a coordinating committee' [P] Again, I'm gonna say [Option C] 'endeavor' haha [chuckle] because it fits.

R: Yeah.

Halle: So, I have three left. I'll just do all of them just to make sure that it is the right word. So, [Line 5] 'or adventure travel. It is a tedious and time-consuming blank [i.e., KWIC position] that requires considerable inconsequential travel over vast expanses of territorial' [IA 'I assume'] So, it's, again, [Option C] 'endeavor' fits, [IA 'it's'] [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] 'time-consuming [Option C] 'endeavor'' [P] So, the sixth line [Line 6] 'devoted to social activity would leave less time for [E] academic [E] blank [i.e., KWIC position] [P] and therefore would likely have a negative impact on academic', again, [Line 6 again immediately, in word, right before the KWIC position, with KWIC inserted] 'academic [Option C] 'endeavor''. That's fit perfectly. And finally, [Line 7] 'being answered, but as is always the case in scientific blank [i.e., KWIC position] more questions have arisen. We haven't yet found evidence for', again, [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] 'scientific [Option C] 'endeavors''. So, I'm gonna go with C 'endeavor' as being the correct answer.

R: So, yeah you can choose

Halle: [Ticking Option C 'endeavor']

R: Uh-huh, what are the key words that help you [P] to decide that it must be [Option C] 'endeavor', not other words?

Halle: Er, well, the words that are [E] close to [E] er [Option C] 'endeavor', so, for example, [Line 6, in word, right before the KWIC position, with KWIC inserted] 'academic [Option C] 'endeavors'', [Line 7, in word, right before the KWIC position, with KWIC inserted] 'scientific [Option C] 'endeavors''. It fits. It wouldn't [Option B] 'difficulty', it wouldn't make sense. [Option D] 'fruitlessness'? No way. Because it's more of an adverb.

R: OK.

Halle: And then [Option A] 'attempt' doesn't fit with s [P] it doesn't make sense in the context, so, first of all, I look at the words that are close to the blank, and then I look at the overall context to make sure it makes sense and fits the whole picture.

R: OK. Perfect.

ConCloze 3: Igor on Item 1 (Target word: *endeavor*)

Igor: OK.

R: Whatever you are thinking, whatever you are er what is going on in your mind, just say it.

Igor: OK, I'm [Item instructions] 'Choose the most appropriate answer.' [P] appropriate answer, then I will have to choose only one, right?

R: Yeah.

Igor: From er four choice.

R: Yes.

Igor: The first one [concordance line] is er [Line 1, in part, left-hand only] 'theoretical detail [P] ultimately undermines the properly interdisciplinary [whispering] nature of the [KWIC position here]'

R: Keep saying.

Igor: [Line 1 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'nature of the [KWIC position here]' hmm [P] this one should be [E] noun [E]. I have to put noun or adjective something like that. [P] should be noun, yes.

R: Uh-huh.

Igor: So, I see the choice [P] [Option A] 'attempt' is not noun anymore, so cut it out. And [Option B] 'difficulty'? [Option B] 'difficulty' is [P] no, cannot. [Option C] 'endeavor' and [P] [Option D] [E] 'fruit fruitlessness' [E]. Should be [IA 'A'], C, and D, I'm not sure, me see let me see. I dunno the the meaning as well, because it's er

R: No worry, no worry. Whether your answer is right or wrong.

Igor: OK, ok, I just just choose C [Option C 'endeavor'] and D [Option D 'fruitlessness'], so [Line 2, in phrase, from right before the KWIC position, with KWIC inserted] 'human [E] [Option C] 'endure [E] [endeavor?]' [P] as music' [P] [Line 2 again immediately, in part, with KWIC inserted] 'their world, but that education in such primary [E] human [E] [P] hmm [Option C] 'endeavor as' [P] [IA]

R: Keep saying.

Igor: Yeah, [Note that Line 3 appears to have been skipped here.] [Line 4, in part, left-hand only, with KWIC inserted] 'and [P] coordination across divisions of the college must [P] occur. This [P] [Option C] 'en endeavor'', right? [Option C] 'endeavor', I like [Option C] 'endeavor', I don't know what's [P]

reasonable is. [P] Ah, this should be the meaning, ok. [Line 4 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'This [KWIC position here] is a campuswide effort that the beginning [P] with a cor coordinating committee.' [P] And [Line 5, in part] 'or adventure travel. It is a tedious and [E] time-consume [E] [in fact 'time-consuming'] [P] [KWIC position here] that requires constr' psst [P] [Line 6, in part, left-hand only] '[de]vote[d to] social ['activity' not verbalized] leave less [IA 'minutes'] for academic [KWIC position here]' [P] I, OK, I I decide to choose the C [Option C 'endeavor']. I dunno the meaning but it should [P] make sense more than another choice [i.e., other options] [P]

R: Can [P] You you can tick that.

Igor: I choose it?

R: Yeah.

Igor: OK, [IA] [Option B] 'difficulty' [P] let me er hmm [Option B] 'difficulty' [P] [Line 4, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'difficulty' is' [P] psst [P] hmm [Line 2, in word, right before the KWIC position] '[IA 'primary'] [KWIC position here]'

R: Keep saying.

Igor: OK, yeah, yeah. Just guess this one. I I don't know much about vocab.

R: Moment, er, so you decided to choose Choice C [Option C 'endeavor']. Why?

Igor: Yeah, in case in case I don't know the meaning. I just see what kind of the the word like noun, adjective, and adverb something like that. Because I don't know much about vocab.

R: So, you mean you do not know the meaning of Choice C [Option C 'endeavor']?

Igor: Some of them. Some of them I know, yeah. But [P] I don't know exactly the meaning, but I see should be the why I guess. I dunno.

R: Why, why do you choose the choice whose meaning you do not know?

Igor: Hmm, [Option A] 'attempt' I I know it, but it's not 'attempt'. [P]

R: Why?

Igor: This one, this one should be word [noun??]. [Option A] 'attempt' is, I think, is verb, right? Like something you need to acting something, do something, try try to do something, [Option A] 'attempt'. Choose. So, I try to cut some of them out [Obviously based on his understanding that Option A 'attempt' can only be a verb.]

R: Uh-huh.

Igor: And guess er [Option B] 'difficult[y]'? Wait, OK, [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] 'the [Option B] 'difficulty' and results' is not make sense, it's not match together. I dunno, so cut of them.

R: Uh-huh.

Igor: Maybe, [Option B] 'difficult[y]' is not appropriate some of they. So, I I don't know this one [Potentially, the selected option, Option C 'endeavor'] correct or not, but I choose the [P] the one that I think is wrong

R: Out

Igor: First, yes.

R: And why, what about Choice D [Option D 'fruitlessness']?

Igor: [Option D] 'fruitless' I'm I'm, 'fruitlessness'

R: So, do you think it is also

Igor: 'Fruit' is like something for eat to [P] but [Option D] 'fruitlessness', it should be [P] just a little bit, I guess the meaning, but when I see hmm psst about something for eat something like that, I don't see it. I think, should be C, in my opinion.

R: O, alright.

Igor: I dunno. Maybe it's wrong. I just guess. This is the way to guess.

R: Yeah. No problem.

Igor: You have to cut something that is [P] 100% wrong [E] out [E], and

R: Hmm.

Igor: OK your chance to get the right answer that higher. From four choice you get three. That's mean the percent to correct is increase.

R: Uh-huh.

Igor: Erm, I don't know. OK.

ConCloze 4: James on Item 1 (Target word: *endeavor*)

James: [IA]

R: Speak and think at the same time.

James: [IA 'you want me to speak this?']

R: Whatever, whatever you are thinking, whatever you are reading, just say it. No worries.

James: [Line 1] 'theoretical detail ultimately undermines the [P] properly interdisciplinary nature of the blank [i.e., KWIC position] and the results in a dominance of the social sciences in' [P] [item instructions] 'All the lines above miss the same word. Which of the following should be [IA 'that word']?' [IA 'in' [from Line 1]] [Line 2] 'their world but [IA 'that'] education in such a primary [E] human [E] [P] er blank [i.e., KWIC position] as er music should be universal for our children in their' yeah [Line 3] 'logical to go with your strengths [IA 'and to'] structure [P] the blank [i.e., KWIC position] in the way [in fact 'in such a way'] ['as' not verbalized] to guarantee success by anticipating' [P] [Line 4] 'and coordina[tion] across divisions of the college must occur. This blank [i.e., KWIC position] is a campuswide effort and the begin that ['with a' not verbalized] coordinating committee' [P] Oooh [IA] [Line 5] 'or adventure travel.' To same blank oh

R: Sorry?

James: Interconnected?

R: No, no,

James: No?

R: The seven lines are separate.

James: O, OK.

R: They are not from the same place.

James: [Line 5 continued] 'this is [in fact 'It is a'] tedious and time-consuming blank [i.e., KWIC position] [P] [E] and [E] hmm, that requires considerable inconsequential travel over vast expanses of [IA 'territory']' [Line 6] 'devoted to social activity would leave less time for academic [P] er blank [i.e., KWIC position] and therefore would like to [in fact 'likely'] have a negative impact on academic' [P] [Line 7] 'being answered but [P] as is always the case with in scientific blank [i.e., KWIC position] more questions have arisen. We haven't yet found evidence [E] for [E]' [Item instructions again] 'All the lines above miss the same word. Which of the following should [P] be that word?' [P] [sigh] OK, [P]

R: Keep saying, whatever you are thinking.

James: [IA 'travel' from Line 5?] [P] The first one is [Option A] 'attempt' er [IA 'and therefore' from Line 6?] just apply the haha [chuckle] the that key word to the all those seven words [i.e., concordance lines?] and [P] I have to think if its er make any sense. [P]

R: Yeah.

James: [Option C] 'endeavor' [Line 2, in part, with KWIC inserted] '[IA 'education] in such a [E] primary human [E] [Option A] 'attempt' as a music' [P] yeah, hmm, [P] [Option C] 'endeavor' [P] and [E] [Option D] 'fruitlessness' [E] [P] hmm, oh that, hahaha [chuckle]

R: No worries, yeah. Keep saying.

James: [Line 4, in word, left-hand only] 'coordinating' er [Line 6, in part, left-hand only] 'devoted to social activity would leave less time for academic [KWIC position here]' I think er the most suitable one is [Option A] 'attempt'

R: Uh-huh.

James: I [P] and I can't substitute other so so I'm, honestly, I don't know the meaning of that one.

R: Which one? Choice C [Option C 'endeavor']?

James: Yeah. And I, last one, D [Option D 'fruitlessness'].

R: O, o, Choice D [Option D 'fruitlessness'], uh-huh

James: Yeah, yeah, [P] and that make some er make sense if I apply the [Option A] 'attempt' to all this because er [P] it's er [P] it's er I mean [P] I think it's most suitable one er because

R: Uh-huh. So, from these seven lines [P] er what are the key words in these seven lines that [E] help [E] you to come to this [P] conclusion?

James: Because in [Line 2, in part, left-hand only] 'their world but that education in such a [E] primary [P] human [E] [KWIC position here]' that couldn't be er [Option B] 'difficulty' [P] and [Option C] 'endeavor'. Should be [Option A] 'attempt'. That's my, [IA 'something prime'].

R: You mean 'primary human' [from Line 2] and then you

James: [Option A] 'attempt'

R: think it should

James: [Option A] 'attempt'

R: Be [Option A] 'attempt', uh-huh. Any other words in other lines?

James: Er, and the, [Line 5, in part, left-hand only] 'or adventure travel. It is ['a' not verbalized] tedious and [E] time-consuming [E] [KWIC position here]' so so [Option A] 'attempt' might the time and doing doing some sense [Line 7, in part, with KWIC inserted] 'being answered but as is always the case in scientific [Option A] 'attempts''. It's not [Option B] 'difficulty', [Line 7 continued] '[KWIC position here] more questions have arisen.' [P] That's mean that the the the [Line 7 again immediately, in phrase, right after the KWIC position] '[Option B] 'difficulty', the [no 'the' here in the real concordance line] more questions have arisen', er not. But but when I conducting scientific [Option A] 'attempts', the [IA] more questions arise. [Note the adoption of the words from Line 7 into the explanation of the answer.]

R: Uh-huh.

James: Actually, [E] some [E] lines I can substitute the [Option B] 'difficulty' as well but er, yeah, but it's not more suitable for the others [i.e., options]

R: Uh-huh.

James: But I feel, if I er substitute [Option A] 'attempt', I think it's more suitable

R: Alright, OK.

[...]

R: Yeah, yeah. And you assume what?

James: I assume this one I don't have [IA 'meaning'] for me.

R: It is ok. That that is normal. OK, next please.

James: [Can't go to the next page.]

R: You can tick it.

ConCloze 4: Klavier on Item 1 (Target word: *endeavor*)

R: And speak and the same time.

Klavier: [P]

R: Keep saying.

Klavier: Oh, [IA 'I have to read things like that']

R: Up up to you. Yeah,

Klavier: Of course.

R: If you are reading, of course, you can read it out loud.

Klavier: I was just thinking, yeah. Erm [Line 1] 'theoretical details ultimately undermines the properly inst in[ter]disciplinary erm nature of the [P] [KWIC position here]' [P]

R: Keep saying.

Klavier: [Line 1 continued] And '[KWIC position here] and results in in a dominance of the social sciences [P] er in' [Line 2] 'their world but that education in such a primary human [P] [KWIC position here] as music should be universal for our children in their' [P] [IA] [Line 3] 'logical to go with your strengths and to structure the [P] [KWIC position here] in such a way as to guarantee success by anticipating' [Line 4, in part, left-hand only] 'and coordination across divisions of the college must occur. [P] This [KWIC position here]' I think it it's [Option A] 'attempt'

R: Uh-huh.

Klavier: [P]

R: Why?

Klavier: [P]

R: Could you please explain why do you think it should be [Option A] 'attempt', not the other options?

Klavier: Erm, [P] I think it kind of fits there. [P]

R: What do you mean, 'fit'? With what?

Klavier: Erm, for example, [Line 4 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'This [Option A] 'attempt' is a campuswide effort' [P] erm

R: Uh-huh.

Klavier: So, an 'effort' [from Line 4] can be an attempt [Option A 'attempt'].

R: Uh-huh.

Klavier: Erm, [P] and the [Line 2, in word, right before the KWIC position, with KWIC inserted] 'human [Option A] 'attempt''

R: Uh-huh.

Klavier: Erm [P] I think that's correct. Is that correct?

R: No, I cannot say, but, yes, you can tick it.

Klavier: O, OK.

ConCloze 4: Lulu on Item 1 (Target word: *endeavor*)

Lulu: So, [Instructions] ‘You may need to maxi[mize]’ oh, OK. ‘Choose the most appropriate answer.’ Right. Just the same task [as the sample item].

R: Yeah.

Lulu: [Line 1] ‘theoretical detail ultimately undermines the properly interdisciplinary nature of the blah blah [i.e., KWIC position] and results in a dominance of the social sciences in’ [P] [Line 1 again immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘nature of the [P] [Option A] [E] ‘attempt’ [E]’, [Line 1 again immediately, in word, right before the KWIC position, with KWIC inserted] ‘the [Option B] ‘difficulty’’, [Option C] ‘endeavor’, and er [Option D] [E] ‘fruitlessness’ [E] [P] Hmm, psst, from this first sentence I think maybe [P] [Line 1 again immediately, in part, right-hand only] ‘[KWIC position here] [IA ‘and’] results in a dominance of the social sciences’ I think maybe [Option A] ‘attempt’? [P] Or [Option C] ‘en endeavor’. I don’t know, but I have to go to the next one. [P] Number two, [Line 2] ‘their world, but that education in [E] such a primary human [E] [P] [KWIC position here] as music should be universal for our children in their’. Er, in this sentence, I think er, psst, you should be [Option C] ‘endeavor’ [P] er to indicate if something hmm [P] er [Line 2 again immediately, in part, left-hand only] ‘but that education in such a primary human [KWIC position here]’ It’s something that you do you you take part er so I think it should be [Option C] ‘endeavor’ er [Line 3] ‘logical to go with your strengths and to structure [P] [KWIC position here] in st such a way as to guarantee success by anticipating’ hmmm, psst, yes, I I think I’m, I erm, I’m going to make up my mind, and say I am going for C [Option C ‘endeavor’] for this er for this for this question. Erm and I I I don’t think I want to read the following of er of of questions. Hmm.

R: Yeah, OK. You can choose it [i.e., tick the circle of Option C ‘endeavor’] and then ‘Next’.

Lulu: O, right.

R: OK.

ConCloze 2: Aaron on Item 3 (Target word: *hypothesize*)

Aaron: Hahaha [laughing]

R: I like the way you you read the word and then you say which word you are reading.

Aaron: OK.

R: OK.

Aaron: Hmm haha I I [P] I feel a little scared of this kind of question haha [laughing] [P]

R: Just say it, whatever you are thinking or reading.

Aaron: Yeah, yeah. [P] Erm [Line 1, in part, left-hand only] ‘It is not [E] unreasonable [P] to [E] [KWIC position here]’ [Line 1 again immediately, in part, left-hand only] ‘It is not unreasonable to [KWIC position here]’ [P] [Line 2] ‘knowledge will lead us to [P] [KWIC position here] [IA, murmuring] different authorial and narrative audiences’ [P] er [IA]

R: Just say it.

Aaron: Yeah, I’m I’m still reading er [Line 3, only the last word heard] ‘consumption’ [Line 4, in part, left-hand only] ‘reading of the literature. They [E] erm [E] [KWIC position here]’ [P] [Line 3, in phrase, right before the KWIC position] ‘one could also [P] [KWIC position here]’ [Line 3 again immediately, in phrase, right before the KWIC position] ‘one could [P] [KWIC position here] [IA]’ [Line 3 again

immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘one could also [P] [E] erm [E] [P] [Option C] ‘suggest’’. Er. [P] When I read the this question [Line 5, in part] ‘[Note that the word ‘1992).’ at the beginning of the concordance line seems to have been ignored altogether.] In other words, we [KWIC position here]’ [P] and then ‘the teenager sexual behavior would be’. I think in this sentence [Option A] ‘formulate’ is not appropriate and

R: Uh-huh

Aaron: And er, mind if I [IA] erm I didn’t notice. I need changes in the previous questions. So, I think the, for me the answer will be in [Option B] ‘hypothesis’ [sic] and [Option C] ‘suggest’. But if I use [Option B] hypothesis [sic] in the this question [i.e., Line 5] and the fourth question [i.e., Line 4] [P] I think if I use hypothesis [i.e., Option B ‘hypothesize’] I need to follow by the the qua the my [Option B] ‘hypothesis’ [sic] after the word [Option B] ‘hypothesis’ [sic] [IA] for example my [Option B] ‘hypothesis’ [sic] er [P] someone uh will read the the survey. They read the poll it’s er a full sentence. But in the fourth question [i.e., Line 4], it’s it’s right they [P] worked and found so I think it’s er [Option B] ‘hypothesis’ [sic] would be not be appropriate for this [i.e., Line 4]. So, [P] I think it’s er su [Option C] ‘suggest’. And [Option C] ‘suggest’ is also is er er frequently used the word in the in in my language use haha [laughing].

R: Alright, ok, er. Which word do you pay particular attention to or do you read the whole er back part?

Aaron: Er, I think I read one sentence after another but when I read the fifth question fifth line [i.e., Line 5] I think [P]

R: The fifth line?

Aaron: Yeah, the fifth line [i.e., Line 5] [P] helps me [P] maybe [Option C] ‘suggest’ is appropriate and I haven’t read the last two sentences. [P]

R: Alright. Thank you so much. The next one please.

ConCloze 2: Björn on Item 3 (Target word: *hypothesize*)

R: I want to know the way you think. I don’t care about the passage.

Björn: OK, [Line 1] ‘It’s not unreasonable to er [P] [KWIC position here] a heightened sense of [IA ‘territoriality’]’ OK, and [Line 2] ‘language [in fact ‘knowledge’] will lead us to [P] [KWIC position here] different [IA ‘authorial’] and narrative audiences’ [Line 2 again immediately, in phrase, right before the KWIC position] ‘will lead us to [P] [KWIC position here] [IA]’ [P] erm

R: Keep saying.

Björn: I think [P] this sentence er they are all from the research [P] article?

R: Yeah.

Björn: And [P] so, [P]

R: They are all from academic area.

Björn: [IA] So, answer A [Option A ‘formulate’] or B [Option B ‘hypothesize’] will have higher possibility for the answer, [P] for er [Line 6, with KWIC inserted] ‘happier than single people and [Option B] ‘hypothesis? hypothesis’ [sic] marriage may draw persons’ So they give [IA ‘a fact or a fact’] or sometime things they have found. And they give some erm [Option B] ‘hypothesis hypothesis’ [sic] er ok so. The answer B? [i.e., Option B] ‘hypothesis’ [sic]. [Line 7, with KWIC inserted] ‘behavior. Notably parental smoking is [P] [IA [Option B] ‘hypothesis’ [sic]] to [P] demonstrate pro-smoking norms and’ OK. It’s [Option B] ‘hypothesis?, hypothesis’ [sic]. OK, I choose the answer B [Option B ‘hypothesize’].

R: OK.

Björn: Sounds make sense [IA] [sigh]

ConCloze 2: Claire on Item 3 (Target word: *hypothesize*)

Claire: [Line 1] 'It is not unreasonable [E] to [E] blank [i.e., KWIC position] a heightened er sense of [IA 'territoriality']' [P] [Line 2, in part, left-hand only, with KWIC inserted] 'knowledge will lead us to hmm [IA] narrative audiences', [Line 2 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'lead us to [Option A] 'formulate'' [Line 2 again immediately, in word, right before the KWIC position, with KWIC inserted] 'to [Option B] 'hypothesize'', [Line 2, in word, right before the KWIC position, with KWIC inserted] 'to [Option C] 'suggest'', erm [Option D] 'verify'?' [Line 3, in part] [IA] 'Alternatively, one could also hmm blank [i.e., KWIC position] that adolescents [P] with high consumption' [Line 4] 'reading ['of' not verbalized] the literature hmm they [KWIC position here] [IA 'and'] found three different clusters' [Line 5, in part, left-hand only, with KWIC inserted] 'In other words we [Option C] 'suggest suggest'?' [IA]

R: Say it, clearly.

Claire: [IA] [Line 1, in part, from right before the KWIC position, with KWIC inserted] 'to [Option B] 'hypothesize' [IA 'a heightened sense of']' [P] [Option B] 'hypothesize' [Line 2, in part, from right before the KWIC position, with KWIC inserted] 'will lead us to [Option B] 'hypothesize' different ['authorial and narrative' not verbalized] audiences' [Line 3, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'hypothesize' adolescents' [Line 4, in part, right-hand only, with KWIC inserted] '[Option B] 'hypothesize' hmm ['and' not verbalized] found three distinct clusters' [P] [Line 5, with KWIC inserted] '[IA] other [P] words we er [Option B] 'hypothesize' teenage sexual behavior would be' er [Line 6, in part, left-hand only] 'happier [P] than single persons and [P] [KWIC position here] 'hmm [Line 5, in phrase, the end of the line] 'would be' [Line 6, in word, left-hand only] 'happier than [IA] I think er [P] the [P] clues is the 'found' [from Line 4] [P] yeah

R: Yeah.

Claire: So, [P] before finding something we should [P] er [Option B] 'hypothesize'

R: Uh-huh.

Claire: [Option A] 'formulate', yeah, not [Option C] 'suggest', not [Option D] 'verify'?

R: Uh-huh.

Claire: So, only [IA] we [IA] [Option B] 'hypothesize'. [IA] I think it's number two [i.e., Option B 'hypothesize']

R: OK. Anything else you would like to say?

Claire: No, haha [laughing]

R: OK, let us stop for a while.

ConCloze 2: Dakota on Item 3 (Target word: *hypothesize*)

Dakota: [Line 1, in part, left-hand only] 'It is not unreasonable [E] to to [KWIC position here] [P]'

R: You can scroll.

Dakota: Hmm, [Line 1 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'unreasonable [E] to [E] [Option A] 'formulate' a heightened sense [P] sense of terri reality [in fact 'territoriality']' [Line 2, in part, left-hand only] 'knowledge will lead us [KWIC position here]' [P] I don't think I need to see all.

R: O,

Dakota: I just need er information er [P] some key words, you know, key words [E] before [E] and after the blanket ['blank'??], you know

R: You mean just just [E] short [E],

Dakota: Yeah, short.

R: Just a few words before, and a few words after?

Dakota: Yeah, yeah.

R: O, really.

Dakota: I want to save time [IA] time quite [IA] er pressured, you know. And [Line 2 again immediately, in part, left-hand only] 'knowledge will lead us [E] to [E] [KWIC position here]' [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'lead us [E] to [E] [P] [Option D] 'verify' [Option D] '[E] verify [E]' [Option D] 'verify' difficult [in fact 'different'] authorial' [P] hmm, [P] so [Line 3, in part] 'Alternatively, one could also [P] [Option D] 'verify' that adolescents' No, doesn't sound right.

R: Uh-huh.

Dakota: [Option D] 'verify' with er [Line 4, in part] 'reading of the literature, they [KWIC position here]' [P] O, 'ed' [i.e., the inflectional morpheme in the concordance-prompt line] [Line 4 again immediately, in word, right before the KWIC position, with KWIC inserted, the inflectional morpheme merged] 'they [Option D] formulated' [P] [Line 4 again immediately, in part, from right before the KWIC position, with KWIC inserted, the inflectional morpheme merged] 'they [Option D] formulated [P] ['and' not verbalized] found three [P] distinct clusters' [P] [Distraction: Her phone goes off.] [Line 5, in part, left-hand only] 'In other words, we [KWIC position here]' [P] O, I think my husband is come, so close down here. [IA]

Dakota: [Line 3, in part, left-hand only, with KWIC inserted] '[IA 'reading of the literature'] they er they [Option D] 'verify' [P] [Line 3 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'verified' and' [P] [Line 3 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'they [Option D] 'verified and', [Line 3 again immediately, in word, right before the KWIC position, with KWIC inserted] 'they [Option D] '[E] verified [E]', er no, can't be say that. [Line 5, in part, left-hand only] 'In other words, we [KWIC position here]' [P] hmm [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] 'we [IA [Option B] 'hypothesized']' [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] 'we [Option B] 'hypothesized'' [P] er [Line 5 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'we [Option B] 'hypothesized' teenager sexual behavior would [E] be [E]'. [P] [Line 6, in part, with KWIC inserted] 'happier than [P] sin than single people and and er [P] [Option B] 'hypothesize' that [P] marriage' [P] [Line 5, in word, right before the KWIC position, with KWIC inserted] '[IA 'we'] [Option A] 'formulate'' [P] w ['we'?] [Line 5 again immediately, in part, left-hand only, with KWIC inserted] 'in another word we [Option C] 'suggest' [P] no, no [Line 6, in part, right-hand only, with KWIC inserted] '[Option C] 'suggest', the marriage will draw person' [Line 6 again immediately, in part, right-hand only, with KWIC inserted] '[KWIC position here] [IA] [P] may draw persons' [P] er, [Line 5, in part, left-hand only, with KWIC inserted] 'In another word, we [P] [Option A] 'formulated' [IA]' [P] no. [Option B] 'hypothesized' [P] and [Line 6, in part, right-hand only, with KWIC inserted] '[Option B] 'hypothesized' [P] that marriage may draw persons' [IA] [Line 7, in part, with KWIC inserted] 'Notably, personal [actually 'parental'] smoking [E] is [E] [P] [Option B] 'hypothesized' to' [P] [Line 7 again immediately, in part, from right before the KWIC position] '[IA] smoking [P] [KWIC position here] [a few words not verbalized] pre-smoking [actually 'pro-smoking'] norm and' [P] [Line 7 again immediately, in phrase, right before the KWIC position] 'parental smoking is [P] [KWIC position here]'

R: Say it, whatever you are thinking. No worry.

Dakota: OK. [Line 1, in part, with KWIC inserted] 'It is [E] not [E] unreasonable to [P] [Option B] 'hypothesize' the the ['heightened' not verbalized] sense of' [P] OK, so, [Line 1 again immediately, in part, left-hand only] '[IA 'It's'] [E] not [E] unreasonable [KWIC position here]' [P] I think it's [Option B] 'hypothesize' anyway.

R: OK. Any key words that help you? Could you tell me any key words, in any line, that helps you [P] specially?

Dakota: I think er it's just er when I read here, [Line 5, in part, with KWIC inserted] 'In another words, we [Option B] 'hypothesized' the teenager sexual' and then and then [Line 6, in part, left-hand only] 'happier than single people and [KWIC position here]' then [Line 6 again immediately, in word, left-hand only] 'sing[le] [KWIC position here]' er, maybe, this is, er make a sense.

R: You mean, the way that you have sentence

Dakota: The structure

R: The sentence structure?

Dakota: Yeah, yeah, structure, the sentence.

R: OK.

ConCloze 2: Esther on Item 3 (Target word: *hypothesize*)

Esther: [IA 'and it's the'] [Option A] 'formulate'. [Option B] 'hypothesize'. [Option C] 'suggest', and [Option D] 'verify'. OK. [P] [Line 1] 'It's not unreasonable to dash [i.e., KWIC position] a heightened sense of [E] territoriality [E]', right?

R: Uh-huh.

Esther: [P] Uh-huh, [P] [Line 1 again immediately, in part, left-hand only] 'it's not unreasonable [KWIC position here]', you can even use [Option A] 'formulate' here.

R: Uh-huh.

Esther: Somebody [E] is [E] [IA 'being in a'] research, they can [IA 'do have'] hypothesis [Option B 'hypothesize'] also. I will read a few more.

R: Yeah, yeah.

Esther: [Line 2, with KWIC inserted] 'knowledge will lead us to [P] [Option C] 'suggest' different [P] authorial and narrative audiences'? [P] [Option A] 'formulate' can be used here. [Line 3, in part] 'Alternatively, one could also [P] dash [i.e., KWIC position] the adolescents with high consumption' [P]

R: Yeah, just say it.

Esther: Erm.

R: Whatever you are thinking.

Esther: I I am trying to understand the sentence I [IA] OK, [Line 4, in part] 'reading of the literature they dash [i.e., KWIC position] and found three distinct cluster [P] clusters' OK. So, [P]

R: Whatever you are thinking.

Esther: 'They found' [from Line 4] but they [Line 4 again immediately, in part] 'reading the literature' is also they [IA 've'] given. Here they did something and then they '[E] found [E] three distinct clusters', so [IA 'naturally'] they did not [Option A] 'formulate' it.

R: Uh-huh.

Esther: It was already there, because 'literature' [from Line 4] was given.

R: Uh.

Esther: And so they [P] [Option D] 'verify', no, they [P] psst found, if they found, it means it is already there.

R: Uh-huh. So?

Esther: So, one [P] this word [Option C] ‘suggest’, to me, it’s [Option C] ‘suggest’; they did not [Option A] ‘formulate’ it, because something was already there and they found it. So, [P] [Option C] ‘suggest’, no. [P]

R: Yeah, just say whatever you are thinking

Esther: I I don’t know what to say, because er [P] they [Option C] suggest and found it? [P] What the what did they [Option C] ‘suggest’? [P]

R: What about other lines?

Esther: They’re like they, could that be, OK. [Line 5, in part] ‘In other words we dash [i.e., KWIC position] teenage sexual behavior would be’ [P] [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] ‘we [Option C] ‘suggested’. [P] Or [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] ‘we [Option B] ‘hypothesized’. [P] [Line 6] ‘happier than single people and [E] dash [E] [i.e., KWIC position] that marriage may draw persons’ [P] [Line 6 again immediately, in part, left-hand only] ‘happier than single people [E] and [E] [KWIC position here]’ [P] Now this is, I don’t know, it’s a part of a sentence, so I can’t understand the whole thing [P] erm [P] And it’s no end, no end either. [chuckle] So, I don’t know, like [P]

R: Yeah.

Esther: [P] OK, [Line 7, in part] ‘Notably, parental smoking is dash [i.e., KWIC position] to demonstrate pro-smoking norms’ [P] it’s [Line 7 again immediately, in phrase, right after the KWIC position, with KWIC inserted] ‘[Option C] ‘verified’ to demonstrate pro-smoking’ er [Option C] ‘suggested’ should also. Because er [P] [Line 7 again immediately, in phrase, right before the KWIC position] ‘parental smoking [KWIC position here]’ [P] Because you, like, it’s, I don’t know what’s the next part of the sentence. But prob possibly they’re trying to say that the ‘parents are smoking’ [words from Line 7]

R: Uh.

Esther: It’s it’s [Line 7 again immediately, in part, right after the KWIC position] ‘[KWIC position here] demonstrating pro-smoking norms’? [P] O, frankly, the sentence do not say much [laughing]

R: Yeah, OK.

Esther: That’s why I’m not really sure. Because if it is a research [P] it could [Option C] ‘suggest’ something.

R: Uh-huh, probably.

Esther: And if it is a hypothesis, possibly it’s [Option D] ‘verify’ something?

R: Yeah.

Esther: And if [P] it could be a hypothesis [Option B ‘hypothesize’] also.

R: Yeah, so? In the overall picture?

Esther: But I [P] I think over here, [Option A] ‘formulate’ cannot be. [Option C] ‘sugge[st]’ [P] erm, for this this set of sentences, I don’t want to decide one word. [chuckle]

R: Uh.

Esther: [Option C] ‘suggested’ erm possibly would be a better option, to fit most

R: Uh.

Esther: I don’t know [chuckle]

R: It’s ok. You can can just choose. No wor, no worries whether it is right or wrong. Just, yeah.

Esther: [Finally chooses Option C ‘suggest’]

ConCloze 3: Franz on Item 3 (Target word: *hypothesize*)

R: OK, go ahead. Say whatever you are thinking

Franz: How many questions left?

R: Five altogether. And this is the third, so you have, well, one, two, three to go.

Franz: O, long way to go, indeed.

R: Er, this is half way.

Franz: Haha [laughing] [P] [Line 1, in part, left-hand only] 'It's not unreasonable [E] to [E] [KWIC position here]' [P] OK, it must be verb. [P] Right? Because of [E] 'to' [E] [from Line 1], and when I take a look [P] at the answers, they [E] are all verbs [E]. OK. [P] That's what I'm thinking.

R: Yeah.

Franz: [Line 1 again immediately] 'It's not unreasonable to tut tut tut [i.e., KWIC position] a heightened sense of [E] territoriality [E]' O, psst, this is [P] this sounds very [E] Band-Seven [E] vocabulary for IELTS test. [chuckle] 'territoriality' [from Line 1]

R: OK.

Franz: Erm seven or seven point five. Haha [laughing] [P] [Line 2] 'knowledge will lead us [E] to [E] [KWIC position here] different', huh? psst hmm [P] 'authorial and narrative audiences' [Line 3, in part, left-hand only] 'Alternatively, one could also tut tut tut tut [KWIC position here]' Psst, OK, put it this way, dear, erm, [P] from the context of [E] all [E] questions you provided, you can [E] see [E] erm some clues that er this text erm convey. I think it's about research because I see the [E] word [E] 'unreasonable' [from Line 1] I see 'know', I see word 'knowledge' [from Line 2], right? And [P] psst 'high consumption' [from Line 3] is very big word. And the 'literature' itself [from Line 4], 'literature' here doesn't stand for fiction. It stands for 'previous study', I guess. As a research student, I can know. So, [chuckle] I [P] [Line 4, in part, from right before the KWIC position] 'they tut tut tut [i.e., KWIC position] three distinct clusters' [Line 5, in part] 'In other words, we tut tut tut [i.e., KWIC position] teenage sexual behavior would be' [Line 6] 'happier [IA] than single people haha [laughing] [KWIC position here] that marriage may draw persons' [Line 7] 'behavior. Notably, parental smoking is [KWIC position here] to demonstrate pro-smoking norms' OK. OK. Hmm [P] Hmm

R: Just say it, yeah.

Franz: Well, first of all, I tend to be confused by the words [Option A] 'formulate' and [Option B] 'hypothesize'

R: Yeah.

Franz: Coz it's quite, at some point, synonymous. [E] But [E] if I place [Option A] 'formulate' the word [Option A] 'formulate' into the sentence number seven, it doesn't sound good.

R: OK.

Franz: [Line 7, in part, from right before the KWIC position, with KWIC inserted] 'parental smoking is [P] [Option D] 'formulated' to demonstrate' It it doesn't sound good at all to me. So, according to the [IA 'word'] I told you, 'literature' [from Line 4], [P] 'high consumption' [from Line 3], 'knowledge' [from Line 2] [P] to make it more [E] engaged [E] in the sense of [E] 'research' [E] [the theme he uses?], I rather choose [Option A] 'formulate'. Answer A.

R: Uh-huh, OK. Anything else that you would like to say about this item?

Franz: Hmm, [P]

R: So you, how many lines did you actually use to come to the answer?

Franz: Up to five.

R: Just five?

Franz: Up to five, I guess.

R: You mean number one to number five or

Franz: Num number one, number two, erm, erm, number three, number four, and

R: Seven

Franz: Number seven.

R: Alright.

Franz: Actually, the word [Option D] 'verify', it sound good as well. But if I do place erm the word '[Option D] 'verify' into number seven, it it doesn't sound good, because you have the word 'demonstrate' [from Line 7]. It sounds very awkward, [Line 7, in phrase, right after the KWIC position, with KWIC inserted] '[Option D] 'verify' something to demonstrate'. But it can be [Option B] 'hypothesize' in the sense of research, I guess. I hope I should be correct.

R: Uh, yeah.

ConCloze 3: Gill on Item 3 (Target word: *hypothesize*)

Gill: [Line 1, in part, left-hand only] 'It is er [P] it is un it is not [P] unreasonable to [KWIC position here]' [P] the words are [Option A] 'formulate', [Option B] 'hypothesize', [Option C] 'suggest' [DISTRACTION] [P] [and Option D 'verify'].

R: Yeah, it's OK.

Gill: OK, [Line 1 again immediately, with KWIC inserted] 'It is not unreasonable to [Option A] 'formulate' a heightened sense of territoriality' [P] [Line 2, with KWIC inserted] 'knowledge will lead us to [Option A] 'formulate' different authorial and narrative audiences' [P] [Line 1, in part, left-hand only, with KWIC inserted] 'it is not unreasonable to [Option B] 'hypothesize'' [P] [Line 1 again immediately, in part, left-hand only, with KWIC inserted] 'it is not unreasonable to [Option C] 'suggest'' [P] [Line 2, in part, with KWIC inserted] 'knowledge will lead us to [Option C] 'suggest' different' [P] Psst, [Line 3, in part, with KWIC inserted] 'Aut [i.e., 'Alternatively'] one could also [Option C] 'suggest' that adolescents [IA 'with'] high consumption' [P] [Line 4, in part, with KWIC inserted] 'reading of, reading of literature, reading of literature [E] they [E] [P] [Option C] 'sug [P] gested' [P] and found that [P] distinct, and found three distinct' [Line 5, in part, with KWIC inserted] 'in 1992 [P] in the words [P] we we [Option C] 'suggested' teenage [P] behavior' [Line 6, in part, left-hand only] 'happier than single people and [KWIC position here]' [P] [Line 5, in part, right-hand only] '[KWIC position here] [IA 'the the'] sexual behavior would be' [Line 6, with KWIC inserted] 'happier than single people and [Option C] 'suggested' that marriage may draw [P] may draw [P] persons' [Line 7, with KWIC inserted] 'behavior. Notably [P] parental smoking is [P] [Option C] 'suggested' to demonstrate pro pro-smoking, pro-smoking norms' [P] Yes, I think I will going for [Option C] 'suggest'

R: Yes, uh-huh, so

Gill: Yeah

R: What are the key words that help you specifically?

Gill: Yeah, [Line 1, in part, left-hand only] 'it is un it is unreasonable [KWIC position here]'. Yeah, [Line 1 again immediately, in part, left-hand only] 'it is unreasonable [KWIC position here]', it means that not er [P] not

R: Logical

Gill: Not logical or data driven.

R: Uh-huh.

Gill: Yeah, it is [IA 'unreason']. And so in research if you are going by reason, it is better to use the word [Option C] 'suggest'

R: Uh-huh.

Gill: OK.

R: What about other lines?

Gill: Er [Line 2, in part, left-hand only] 'knowledge will lead us [P] [E] to [E] [KWIC position here]' [Line 1, in part, right-hand only] '[KWIC position here] the [in fact 'a'] heightened sense of territorial[ity]' [Line 2, in word, left-hand only] 'knowledge' [P] here it's the use of [Line 1, in part, right after the KWIC position] '[KWIC position here] heightened sense heightened sense of territorial[ity]' [Line 2, in part, left-hand only, with KWIC inserted] 'knowledge wor will lead us to' yes 'to [Option C] 'suggest''. So I feel that er [P] because it is not empirical, it's [P] by way of reasoning, so it's better to use [Option C] 'suggest'. [P] We could also er [P] we could also er [P] to [Option B] 'hypothesize'. [Line 1, in part, left-hand only, with KWIC inserted] 'It is not unreasonable to [Option B] 'hypothesize'' yeah, we could also use [Option B] 'hypothesize'. But er [P] [Line 1 again immediately, in part, left-hand only, with KWIC inserted] 'it is not unreasonable to [Option B] 'hypothesize''

R: Er moment. Line lines 1 to 7 are not connected.

Gill: Yes, I know, I know, yeah. [P] I know, [Line 1 again immediately, in part, with KWIC inserted] 'It is not re, unreasonable to [Option B] 'hypothesize' a heightened' er of course, [Option B] 'hypothesis' [in fact, 'hypothesize'] is also tentative. [P] But [Option C] 'suggest', [P] let me see, [Line 4, with KWIC inserted] 'reading of the literature, they er [Option C] 'suggested er suggested' and found three distinct clusters' [Line 4 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'they [Option B] 'hypothesized' and found' [Line 4 again immediately, in part, with KWIC inserted] 'In other words we [Option B] 'hypothesized' [P] teenage [P] sexual behavior would be' [Line 6, in part, left-hand only, with KWIC inserted] 'happier happier than single people [KWIC position here]' [Line 6 again immediately, with KWIC inserted] 'happier than single people and [Option B] 'hypothesized' that marriage may draw persons'. [P] I'm tempted I'm tempted to go for [Option B] 'hypothesize'.

R: OK

Gill: Yeah,

R: What, you can go to er Line 7 [IA] so, what are

Gill: [Line 7, with KWIC inserted] 'behavior. Notably parental smoking is [P] [Option B] 'hypothesized' to demonstrate pro-smoking norms'. [P] Yeah, I'm going for [Option B] 'hypothesize'.

R: OK. So, you changed your mind, uh-huh. What are the key words, again, to to make you [P] decide that it it is not [Option C] 'suggest' anymore. It should be [Option B] 'hypothesize'

Gill: Yeah, because er

R: For example?

Gill: For example, er [P] when you read something, [Line 4, in phrase, left-hand only] 'reading [of] the literature', OK, you read literature, you can make [P] [E] a clever guess [E]

R: Uh.

Gill: Uh, So, from 'reading' [from Line 4], you can [Option C] 'suggest' but its but [IA 'what's the'] 'literature' [from Line 4] [IA] is [Option C] 'suggest', it shows that you have read some data. And from there you can [Line 4 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option B] 'hypothesize' three distinct clusters', OK.

R: OK, wow, a real researcher, OK. [chuckle]

Gill: In [Line 5, in part, left-hand only] '1992. In other words we [KWIC position here]' [P] [Line 5 again immediately, in part, left-hand only] '1992 [P] in in other words we [P] we we [KWIC position here]' you you, here here, they are talking about [IA] outcome

R: Uh-huh.

Gill: Yeah, [Line 5 again immediately, in part, right after the KWIC position] '[KWIC position here] teenage sexual behavior'

R: Uh-huh

Gill: So you can [Option B] 'hypothesize' you can make a [E] guess [E]

R: Uh-huh

Gill: Yeah, [Line 6, in part, left-hand only] 'happier happier than single people [KWIC position here]' [P] [IA] [Line 6 again immediately] 'happier than single people yeah [KWIC position here] [P] marriage draw, [P] that marriage draws persons' OK. There is also seeming to be a linkage, you know, between this the phrase the phrase this and the second phrase [left-hand and right-hand parts of the concordance line?].

R: Uh-huh.

Gill: So, yeah [IA 'collectively'] I'll go for [Option B] 'hypothesize'

R: Uh-huh.

Gill: Er, and then here too [Line 7, in part, from right before the KWIC position, with KWIC inserted] 'parental smoking is [Option B] 'hypothesized' to' it's it's [Line 7 again immediately, in phrase, right before the KWIC position] 'parental smoking is [KWIC position here]' [P] er can be linked or can show can lead to, ok, 'demonstrate' [from Line 7] to, OK, 'smoking norms' [from Line 7]. I think [Option B] 'hypothesize', yeah. I think now I'll choose [Option B] 'hypothesize'

R: OK.

ConCloze 3: Halle on Item 3 (Target word: *hypothesize*)

Halle: OK, so the first question. [P] So, first line [Line 1] 'It is not unreasonable to blank [i.e., KWIC position] a heightened sense of territoriality' and the words are [Option A] 'formulate', [Option B] 'hypothesize', [Option C] 'suggest', or [Option D] 'verify'. Again, [chuckle] it becomes more and more difficult. [P] [Line 1 again immediately, in part, left-hand only, with KWIC inserted] 'It is not unreasonable [E] to [E] [P] [Option C] 'suggest' would work, I think, best in this case. [E] [Option D] 'verify' [E] wouldn't [P] really make sense. I think it would be quite hard to [Line 1 again immediately, in part, right-hand only, with KWIC inserted] '[Option D] 'verify' a heightened sense of territoriality'. [P] [Option B] 'hypothesize', again, would work because it's quite similar to [Option C] 'suggest'. And then [Option A] 'formulate', well, that's er more like a formula or something, more precise, more scientific, which you can measure. And [Line 1 again immediately, in phrase, right-hand only] 'sense of territoriality', doesn't seem, so I'm gonna be between [Option C] 'suggest' and [Option B] 'hypothesize', so I'm gonna go with the next one, which is [P] [Line 2] 'knowledge will lead us to blank [i.e., KWIC position] different authorial and narra narrative audiences' [Line 2 again immediately, in part, with KWIC inserted] 'will lead us to' [P] again, '[Option C] 'suggest' different' would work. [P] I think it would work better than [Option B] 'hypothesize' [P] So, I'm gonna go with [Option C] 'suggest' for now. [P] third one, [Line 3] '[E] p [E] [sigh] less than [P] oh point oh oh one hashtag. Alternatively one could also blank [i.e., KWIC position] that adolescents with high consumption' [P] [Option C] 'suggest', I think, [Option C] 'suggest', [Line 3 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'one could also [Option C] 'suggest' adolescents' [P] [sigh] Next, [Line 4] 'reading of the literature [E] they [E] [P] blank [i.e., KWIC position] and found three distinct clusters' [P] again, [Option C] 'suggest' works better because [Option B] 'hypothesize' you do that [P] you come up with an idea and you propose it. And because I have [P] [changed somewhat from Line 4] something coming from the 'literature' [from Line 4]. That [IA 'leads'] me to believe that it is something 'suggested' [adapted from Option C 'suggest'], rather than 'hypothesized' [adapted from Option B 'hypothesized'] [P] I don't think [Option D] 'verify' [P] or the other ones [P] work that well so far. So, I think [Option C] 'suggest' is still right one? Right? I'm gonna verify with the other ones. So, Line 5, [Line 5] '1992. In other words we blank [i.e., KWIC position] teenage sexual [E] behavior [E] would be', again, [Line 5 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'we [Option C] 'suggest' the teenage sexual behavior'. [P] think works best. [P] [Line 6] 'happier than single people and [P] blank [i.e., KWIC position] that marriage may draw persons' and,

again [Option C] ‘suggest’ it [P] it would be [P] just [E] weird [E], so, to use the words like [Option B] ‘hypothesize’ or the [Option A] ‘formulate’ one. Well, talking about ‘marriage’ [from Line 6], I guess. I mean, it seems like informal context, this line [i.e., Line 6], so [P] I’ll go with [Option C] ‘suggest’. And then finally [Line 7] ‘behavior. Notably parental smoking [E] is [E] blank [i.e., KWIC position] to demonstrate pro-smoking norms’ [P] again, [Option C] ‘suggest’ works, so I’m gonna go with [Option C] ‘suggest’, as the overall result.

R: OK. Er moment, please.

Halle: Yes.

R: Er, could you review again, I mean, s say it again, what words help you to decide that it should be [Option C] ‘suggest’?

Halle: I think in this [E] case, [E] [P] it helps me more, the overall context. So, [P]

R: The overall context, rather than

Halle: Yes,

R: Specific words?

Halle: The specific words, and also the tone of the [P] this, for example with Line 6, [Line 6, with KWIC inserted] ‘happier than single people, and [Option C] ‘suggested’ that marriage may draw persons’. It has rather informal [P] tone or [P] either way, it’s not that scientific as the other ones? [P] So, I think [Option C] ‘suggest’ could be used.

R: Yeah.

Halle: And also, because [Option A] ‘formulate’ is for something precisely scientific. It doesn’t work, [Option D] ‘verify’, you have to verify [E] something [E]. [IA ‘that’], again, doesn’t work in the context. So, the one [IA ‘that I think’] is [Option B] ‘hypothesize’. But then when you ‘hypothesize’ something, you come up with the hypothesis [P]

R: Yeah

Halle: And, again, the context. So, for example, Line 4, when it says [Line 4] ‘reading of the literature. [E] They [P] blank [E] [i.e., KWIC position] and found three distinct clusters’ [P] using [Option B] ‘hypothesize’ wouldn’t really work there. So, [Option C] ‘suggest’, it works better, because it fits best, all of them. It’s [P] I think it’s the one [i.e., Option C ‘suggest’] [P] that’s correct, in this case.

R: Alright. Thank you.

ConCloze 3: Igor on Item 3 (Target word: *hypothesize*)

Igor: [Line 1, in part, left-hand only] ‘It is not [P] unreasonable to [KWIC position here]’ [P]

R: Keep saying.

Igor: [Line 1 again immediately, in part, from right before the KWIC position, with KWIC inserted] ‘to [Option A] ‘formulate’ a heightened [E] sense [E] of terri[toria]lity’ [Line 1 again immediately, in part, with KWIC inserted] ‘It is not unreasonable to [P] [Option B] ‘hypothesize’ [P] a [E] heightened [E] sense’ [P]

R: Uh-huh.

Igor: [Line 1 again immediately, in part, from right before the KWIC position, with KWIC inserted] ‘To s [IA ‘answer that’] to [Option C] ‘suggest’

R: Keep saying.

Igor: [Line 1 continued] ‘heightened sense’ [P] [Line 1 again immediately, in part, left-hand only, with KWIC inserted] ‘It is not unreasonable to [Option C] ‘verify’ [IA]’ [IA ‘as well’] [P] [Line 2, in part, with KWIC inserted] ‘knowledge will lead us to [KWIC position here]’ [P]

R: Keep saying.

Igor: [Line 2 continued] 'to [P] [Option D] 'verify' different authorize [in fact 'authorial']' [P] [Line 2 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'to [Option A] 'formulate' different authorize [in fact 'authorial'] and narrative audience' [P] [Line 2 again immediately, in word, left-hand only] 'knowledge' [P] er this one look. [P]

R: Keep saying.

Igor: I I will cut er B [Option B 'hypothesize'] and C [Option C 'suggest'] out now.

R: Alright.

Igor: Just, yeah, in my mind, just need [Option A] 'formulate' and [Option D] 'verify' [P] [Line 3, in phrase, right before the KWIC position] '[E] one [E] could also [KWIC position here]' [P] [Line 3 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'one could also [Option A] 'formulate'' [Line 3 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'one could also [Option A] 'formulate' that adolescent with high consumption' [P] [IA]

R: Keep saying.

Igor: [P] Hmm,

R: Keep saying.

Igor: [IA 'No, it's a bit'] [P] [Note that Line 4 here appears to have been skipped or at least not verbalized.] [Line 5, in part, with KWIC inserted] 'In other words, we [P] [Option A] 'formulate' teenage sexual behavior would be' [P] [Option D] 've verify' [Line 6, in part, left-hand only] 'happier than single [P] people [KWIC position here]' [Line 6 again immediately, in part, left-hand only] 'happier single [P] ['people' not verbalized here] and [KWIC position here]' also [Line 6 immediately, in word, left-hand only] 'happier [IA]'

R: Keep saying.

Igor: And [P]

R: Keep saying please.

Igor: [Line 6 again immediately, in part, right-hand only, with KWIC inserted] '[Option A] 'formulate' that marriage [P] may draw persons' [P] [Line 6 again immediately, in phrase, right-hand only, with KWIC inserted] '[Option D] 'verify [P] verify' that mar[riage]' O, [IA 'sorry'] this one, let me just, psst, [Line 7, with KWIC inserted] 'behavior. Notable [in fact, 'Notably'] paren parental smoking is [KWIC position here]' [P]

R: Keep saying.

Igor: [Line 7 continued] 'is [Option D] 'verify' [P] to demonstrate [P] pro pro-smoking norms' Psst, well, er [P] I I like both of them, [Option A] 'formulate' and [Option D] 'verify', [P] but I know ['don't know?'] which one [chuckle] choose?

R: Uh-huh.

Igor: Psst, [P] [Option C] 'suggest' [Line 7 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'is [Option C] 'suggest' to demonstrate' [P]

R: Keep saying.

Igor: [P] [Line 7 again immediately, in word, left-hand only] 'behavior' psst, no. hahaha [laughing]

R: Well, I would like to know what is going on in your mind,

Igor: Yeah, this one is, this one is

R: So please keep saying.

Igor: Quite difficult for me. It's not familiar. Psst. [IA] [Line 4, in word, left-hand only] 'literature' [Line 4 again immediately, with KWIC inserted] 'reading of ['the' not verbalized] literature, they [P] they

[Option C] 'suggest' and found three [P] ['distinct' not verbalized] clusters' [P] Yeah, this one can use [Option C] 'suggest' as well. [P] [Line 4 again immediately, in part, with KWIC inserted] 'reading of literature review [actually no 'review'], they [Option C] 'suggest' and found three dis[tinct]' [IA] [P] Psst, hmm [Line 4 again immediately, in word, right before the KWIC position, with KWIC inserted] 'they [Option A] 'formulate'' [P] [Line 4 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'they [Option A] 'formulate' and found three' [P] [Line 4 again immediately, in word, right before the KWIC position, with KWIC inserted] 'they [Option D] 'verify' [P]' [P] [DISTRACTION]

R: Keep saying.

Igor: [Sigh] er, all of this is, psst, [P] you can use of them as well, but

R: Yeah,

Igor: It [P] the grammar is no problem, but the meaning [P] [Line 2, in part, left-hand only, with KWIC inserted] 'knowledge will lead us to [P] [Option A] 'formulate'' [P] [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'to [Option C] 'suggest' different authorize [in fact 'authorial']' [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'to [Option A] 'formulate' different authorize [in fact 'authorial']' [Line 2 again immediately, in word, right before the KWIC position, with KWIC inserted] 'to [Option D] 'verify'' [P] psst, I could not decide between [Option A] 'formulate' and [Option D] 'verify'

R: Uh-huh.

Igor: [P] [Line 6, in phrase, left-hand only] 'single people' [Line 6 again immediately, with KWIC inserted] 'happier than single people and [P] [Option D] 'verify' that marriage may draw persons' [P] I choose [Option D] 'verify'

R: OK, why? What are the key words in these sentences that

Igor: I don't know.

R: Help you to

Igor: I just guess.

R: Come to this answer?

Igor: Hmm, between [Option A] 'formulate' and [Option D] 'verify' is [P] can apply [P] to all sentence [P]

R: But

Igor: But er they have some of them is not make sense for me.

R: Which one?

Igor: I think the [Line 6 again immediately, in part, left-hand only] 'hap happier than single people [KWIC position here]' [Line 6 again immediately, in part, left-hand only, with KWIC inserted] '[IA 'happ']ier than single people [E] and [E] [P] [Option D] 'verify''

R: So, you think Line 6 [E] does not [E] [P] what? Let's say, does not

Igor: It's

R: Apply

Igor: To [Option A] 'formulate'

R: To [Option A] 'formulate'

Igor: Yeah. [P] I I think [Option A] 'formulate' for, psst, you do something new [P] or something like that, right? I'm not sure, [Option A] 'formulate'. [P] But [Option D] 'verify' it's mean, psst, [P] you prove something [P] or you get something, I don't know, just guess, OK?

R: Alright. No worries.

ConCloze 4: James on Item 3 (Target word: *hypothesize*)

James: [Line 1] ‘It is unreasonable to [P] blank [i.e., KWIC position] a [E] heightened [E] sense of territoriality’ [Line 2] ‘knowledge will lead us to blank [i.e., KWIC position] [DISTRACTION]

R: Haha [chuckle]

James: What?

R: Go on, go on.

James: [Line 2 continued] ‘[KWIC position here] different authorial and narrative audiences’ [P] [Line 3, in part] ‘Alternatively, one could also blank [i.e., KWIC position] that [P] al adoles [i.e., ‘adolescents’]’ What is that one? I don’t know, [Line 3 continued] ‘with high consumption’ [P] [Line 4] ‘reading of the literature they blank [i.e., KWIC position] and found three distinct clusters’ [Line 5, in part] ‘In another words [in fact ‘in other words’] [E] we [E] [P] blank [i.e., KWIC position] teenage sexual behavior would be’ [Line 6] ‘happier than single people and [P] blank [i.e., KWIC position] marriage may draw [P] persons’ [Line 7] ‘behavior. [P] Notably, parental smoking is [P] blank [i.e., KWIC position] [‘to’ not verbalized] demonstrate pro-smoking norms and’ [P] [Item instructions] ‘All the lines above miss the same word. Which of the following should [P] be the word?’ OK, so [Line 1, with KWIC inserted] ‘It is unreasonable to [P] [Option A] ‘formulate’ a heightened sense of territoriality’ [Line 2, in part, with KWIC inserted] ‘knowledge will lead us to [P] er [P] [Option A] ‘formulate’ [P] hmm different authorial and’ [Line 3, in part, with KWIC inserted] ‘Alternative[ly] one could also [Option A] ‘formulate’ that’ [P] [Line 4, in part, left-hand only, with KWIC inserted] ‘reading [‘of’ not verbalized] the literature they [Option A] ‘formu [P] late’,’ yeah. [P] er [Option C] ‘suggest’, [Option D] ‘verify’. [Line 3, in phrase, right before the KWIC position] ‘one could also [KWIC position here]’ hmm [P] [IA] I think er [P] er most of the time, in [IA ‘other’] word it’s er I think it’s er most suitable one is [Option D] ‘verify’.

R: Uh-huh.

James: And the [P] they talking some facts [P] and the those er those facts could [Option D] ‘verify’ this across. That’s why I choose [Option D] ‘verify’.

R: Uh-huh.

James: Yep.

R: What about other lines?

James: [P] er [P] Are you mean [Option A] ‘formulate’ or the or the

R: I mean, you you mentioned Line 4.

James: Yes.

R: What about the other lines?

James: All all of the [IA] [Line 7, in part, left-hand only] ‘Notably [IA] parental smoking is [KWIC position here]’ It’s kind of, er it came [IA ‘with the’] verification, [Line 7 again immediately, in part, right-hand only] ‘[KWIC position here] to demonstrate [IA] pro-smoking’ yeah this this statement [Option D] ‘verify’ the er fact, the second statement. [P]

R: Uh-huh.

James: That’s why I er [Option D] ‘verify’. [P] er anyway, I think it’s four of them equally give me some meaning but er, according to my personal experience, I think [Option D] ‘verify’ most of [IA ‘suitable’]. If some other words but I don’t think it’s [IA ‘give me’] some sense.

R: Uh-huh. OK, anything

James: Er [IA ‘other’] hard er hardline [IA]

R: Uh-huh.

James: Anyway the [Line 4, in part, left-hand only] ‘reading [‘of’ not verbalized] the literature [IA] [KWIC position here]’ those are the literature they [Option D] ‘verify’

R: Uh-huh.

James: [Line 4 again immediately, in phrase, left-hand only] ‘reading [‘of’ not verbalized] the literature’, I think, not er, you cannot choose the [Option B] ‘hypothesize’. So, if I choose [Option D] ‘verify’ [IA]

R: OK, yeah.

ConCloze 4: Klavier on Item 3 (Target word: *hypothesize*)

R: So that I know what you are thinking about.

Klavier: Erm [Line 1] ‘It is unreasonable to [KWIC position here] [P] heightened sense of territoriality’ [P] [Line 2] ‘knowledge will lead us to [KWIC position here] [P] different authorial and narrative [E] audiences [E]’ [P] [Line 3, in part] ‘Alternatively one could also [KWIC position here] [P] adolescents with high consumption’ [P] [Line 4, in part, left-hand only] ‘reading of the lit literature, they [P] er [KWIC position here]’ [P]

R: Keep saying, whatever you are thinking.

Klavier: I [P] I’m just try to like reread the answers [probably referring to the concordance lines]. And the [P] I’m thinking about [Option C] ‘suggesting er suggest’ er [P] I think I think it’s er [Option C] ‘suggested’ I was just reading the rest of the erm [P] sentences.

R: Uh-huh.

Klavier: Erm, yeah, I think it’s [Option C] ‘suggested’, because it’s kind of fits all of the sentences. [IA]

R: Uh-huh. What are [P] the key words

Klavier: Yes.

R: That help you that you think that it must be [Option C] ‘suggest’, not other choices.

Klavier: Well, erm, the other choices that I kind of erm yeah link to [P] the something else, as in line [P] like specific things like [Option A] ‘formulate’, kind of like you can formulate erm a ‘hypothesis’ [Note that Option B is ‘hypothesize’], for example,

R: Yeah.

Klavier: And the [Option B] ‘hypothesize’ erm. [P] [Option C] ‘suggesting’ is kind of [P] a general words er a general term for [P] something. And [Option D] ‘verify’, it’s like you have to verify something. Erm [P] psst, erm I think it’s [Option C] ‘suggesting’. It’s [P] For a [P] it’s erm it’s scientific research.

R: Uh-huh.

Klavier: And well, they they normally just [Option C] ‘suggest’ and then they [Option D] ‘verify’.

R: Uh-huh.

Klavier: Erm, yeah.

R: Alright, ok.

ConCloze 4: Lulu on Item 3 (Target word: *hypothesize*)

Lulu: Next one, [P] [Line 1] ‘It is [E] not [E] unreasonable [E] to [E] [KWIC position here] [P] a [E] heightened [E] sense of er [E] territoriality [E]’ [P] [Line 2] ‘knowledge will [IA] lead us to [P] [KWIC position here] different authorial and narrative audiences’ [P] [Option A] ‘formulate’, er [Option B] ‘hypothesize’, [Option C] ‘suggest’, [Option D] ‘verify’ [P] er [P] Well, I feel like none of this [options] fits the sentences, haha [laughing]

R: [chuckle]

Lulu: [Line 1, in part, left-hand only] 'It is not unreasonable to er [KWIC position here]' [P] [Line 1 again immediately, in phrase, right before the KWIC position] 'unreasonable to [P] [KWIC position here] [IA]' Hmm, [Line 3, in part] 'Alternatively one could also [P] [KWIC position here] adolescents with high consumption' [Line 3 again immediately, in part, left-hand only, with KWIC inserted] 'one [E] could also [E] [P] [Option C] 'suggest'?' [P] '[Option C] 'suggest'? I think for this one erm [P] it's either [Option C] 'suggest' or [P] [Option B] 'hypothesize'. [P] [IA] [Line 4] 'reading of the literature they [KWIC position here] found three distinct clusters' [P] Hmm, [P]

R: Keep saying.

Lulu: Er, I don't know. I feel like I'm I'm being confused now. Haha [laughing] [Line 4 again immediately, with KWIC inserted] 'reading of the literature they [P] [KWIC position here]' Like I cannot use [Option C] 'suggested [P] suggested' and found three distinct clusters' hmm, no [P] [Line 4 again immediately, in part, right-hand only, with KWIC inserted] '[Option B] 'hypothesized' [P] and found [E] three distinct [E] clusters' [P] [Line 5, in part] 'In other words we [KWIC position here] teenage behavior [IA 'would be']' [Line 5 again immediately, in word, right before the KWIC position, with KWIC inserted] 'we [Option B] 'hypothesized'' [P] This 'would' [from Line 5, in the subordinate clause], isn't it? 'would' should be some kind of er [P] shouldn't be [Option D] 'verify'. If you [Option D] 'verify', it's not 'would', and it's not [Option A] 'formulate'

R: Uh-huh.

Lulu: [Option B] 'hypothesize' [IA] [P] [Line 6] 'happier than single people and [P] [KWIC position here] [that not verbalized] marriage may draw persons' [P] [Line 7, in part, left-hand only] 'Notably, parental smoking [E] is [E] [KWIC position here] [P]' er [P] I think the last one [i.e., Line 7] tells me that it should be [Option B] 'hypothesized'. [P] Hmm, [P] yes, I think I'm going to go for [Option B] 'hypothesize' for this one.

R: Why? What is special about the last one?

Lulu: [P] Er [sigh] [Line 7 again immediately, in part, right-hand only, with KWIC inserted] '[Option B] 'hypothesized' to demonstrate [IA 'a pro-smoking norms']' hmm [P] because er [P] I don't think you can use [Option C] 'suggest' here in this context, because 'parental smoking' [from Line 7] erm [P] you normally use [Option C] 'suggest' with an idea. You [Option C] 'suggest' something to somebody. And how they [IA] [IA 'to the others'] and then you can [Option C] 'suggest' to someone, or it's not a person that you can make a suggestion to, [P] so [P] erm [E] but [E] it it can be the the object of a 'hypothesis' [Option B?] erm [P] so this's why I feel like [IA 'I would go'] for this one. [P] Psst, I can I can go for the [P] hmm [Now Lulu clicks to the next item].

ConCloze 2: Aaron on Item 4 (Target word: *recreational*)

Aaron: I I already have no feeling [P]

R: Just say it, whatever comes to your mind.

Aaron: A little boring of this kind of question. Haha [laughing]

R: Sorry.

Aaron: [Line 1, in part, left-hand only] 'commitments in their lives such as family, employment, and [E] other [E] [KWIC position here]' [P] erm [P] [Line 2, in part, left-hand only] 'If their use of drugs is [E] often [E] [KWIC position here]' [P] hmm, psst [P] [Line 3, in part, left-hand only] 'appear to yield different results from comparisons between [E] competitive [E] and [E] and [E] and [KWIC position here]' psst [P] This question is [E] is [E] is difficult for me, because [P] I'm not quite familiar with the use of erm this word. [P]

R: You mean [Option A] 'avian'?

Aaron: Yeah.

R: Uh-huh, uh-huh.

Aaron: I I seldom use this word [i.e., Option A 'avian'], so I'm not quite sure [P] the paraphrase of this kind of words [the meaning? The synonym?], [E] which [E]. I'm not I I'm [P] I read this word before, but I didn't specially look after the dictionary to know the meaning and the special use of this word. So, erm [P]

R: What about other choices?

Aaron: Other choices? Hmm [P] I have to say, the four choices are not er quite familiar to me, compared to previous er choices.

R: Uh-huh.

Aaron: And, hmm [P] but this word [i.e., Option A 'avian'] is is a little strange [P] to me. And the other three [i.e., Option B 'fluffy', Option C 'mechanical', and Option D 'recreational'] is [P] normal, but I seldom use these four choices. [P] I seldom use them.

R: So, which one?

Aaron: Which one?

R: Which choice?

Aaron: Which choice? Hmm, [Note that Line 4 has now been skipped at this point] [Line 5, in phrase, right before the KWIC position] 'the use of 3 or more [KWIC position here]' [P] [Line 5 again immediately, in phrase, right before the KWIC position] '[E] 3 or more [E] [KWIC position here]' [P] [IA] er, psst [P] hmm, psst [P] I don't know. I don't know. [P] er [Line 2, in phrase, right before the KWIC position] 'is often [KWIC position here]'. Er when I read the second sentence [i.e., Line 2], it's [IA 'quite'] [Line 2, in phrase, right after the KWIC position] '[KWIC position here] associated with leisure activities'. So, [P] 'leisure activities' [from Line 2], I think, is already about the recreational [word from Option D 'recreational'] and entertainment things. Er, [P] so if I still use [Option C] 'recreational' in the second sentence, it's a a little repeatable, so er, [P] I want to exclude the [Option C] 'recreational' [P] hmm

R: But the you have seven sentences [P]

Aaron: Yes, yes, [P] erm

R: All of them just want more [P] one word.

Aaron: Erm [IA]

R: Just say it, whatever

Aaron: [Line 5, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'Mechanical' drugs' is is I sel I seldom see this kind of paraphrase ['phrase?'], [Line 5 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'mechanical' drugs'. It's it's it's not familiar to me. [Line 5, only one word, right after the KWIC position, with KWIC inserted] '[Option A] 'avian' drugs'? I'm not, not, I cannot the meaning! haha [laughing]

R: It's alright. It's just a test.

Aaron: Hmm [P] I don't know. I really don't know. [P] [Line 5, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'fluf [P] fluffy' drugs'? Psst, [P] no no no, I don't know. [P] I don't know I [P] I choose the I choose [Option A] 'avian'.

R: Why?

Aaron: I [P] I don't know the I don't know which one to choose. But the other three [i.e., options other than Option A 'avian'] I think the [P] when I put the [IA 'other'] three words [i.e., options other than Option A 'avian'] in the sentences and I read them, I feel a little weird. [P] And I didn't know the s the use and the meaning of

R: [Option A] 'Avian'.

Aaron: Yeah, I I [E] may [E] know the meaning, but I I still [P]

R: Forget

Aaron: Yeah, yeah, yeah. I seldom use the word.

R: That's alright. OK.

Aaron: [Finally decides on Option A 'avian']

ConCloze 2: Björn on Item 4 (Target word: *recreational*)

Björn: And [sigh]

R: [chuckle] Just say whatever comes to your mind

Björn: [Line 1] 'commitments in their lives such as family, employments and other [P] [KWIC position here] erm, activities and hobbies that capture their interest. The college players' [P] It's about er. [P] [Line 1 again immediately, in part, left-hand only] 'commitments in their lives? [P] Such as family, [P] [IA 'employment'] [KWIC position here]' must be adjective. [IA] [P] [Line 2] 'frequently [in fact 'frequent'] use. If their [P] their use of these drugs is often [P] [IA] [KWIC position here]' [Line 2 again immediately, in part, from right before the KWIC position] 'often [IA] [KWIC position here] and associated with leisure activities, [IA 'then it is likely that']' [Line 3, in word, left-hand only] 'appear O [IA 'what's come to my mind is erm']' [Line 3 again immediately, in part, left-hand only] 'scientific [IA] from comparative [in fact 'comparisons between competitive'] and [KWIC position here]' [P] [IA] [IA 'Sentence 3?'] [P] [IA 'everyone have the same'] er [IA 'nature of the words'?] so [P] [Line 3 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'comparative [in fact 'competitive'] and [Option C] 'mechanical' [P] athlete athletes' [P] [Line 4, in part] 'the most most marinas. A large number of respondents [P] al also commented that [P] [KWIC position here] [IA] boaters are [IA]' [P] erm [P] hmm hmm

R: Say it.

Björn: Erm [chuckle] Some sometime I'm thinking I erm [P] can't speak.

R: Right, I want to know

Björn: Yeah, yeah, yeah.

R: I want to know the way you think.

Björn: [IA] yeah, I know. [P]

R: Yeah.

Björn: Erm for what I am thinking, OK.

R: It's the reason why I need to interview you, yeah.

Björn: Erm erm, yeah [IA] [P] [Line 7, in part, left-hand only, with KWIC inserted] '[IA 'to the United States'] where most bicycles sold are [P] for [P] [Option A] 'avian avian''? So, I don't know what's this word mean, but erm [P] Look like [Option A] 'avian' also. [P] [Line 6, in phrase, right after the KWIC position, with KWIC inserted] '[Option A] 'avian' needs, so improvement' [Line 7, in word, right after the KWIC position, with KWIC inserted] '[Option A] 'avian' use' erm [P] [Line 3, in word, right after the KWIC position, with KWIC inserted] '[Option A] 'avian' athletes' erm [P] [Option B] 'fluffy', [Option C] 'mechanical', [Option D] 'recreational'. [IA], no. [P] [Line 7, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'recreational' use', [P] [Line 7, in part, right after the KWIC position, with KWIC inserted] 'recreational use most of the [IA] bicycles sold [IA]' [P] er for [IA Option C 'mechanical'] [Line 6, in part, left-hand only] 'how [IA] the library [IA] users' educational? [IA] [KWIC position here]' [P] er [P] Shall I? Didn't know some of the words' meaning, so

R: It's OK.

Björn: [P] I choose the most make sense word, [Option C] ‘mechanical’? [Line 7, in word, right after the KWIC position, with KWIC inserted] ‘[Option C] ‘mechanical’ use’, OK.

R: OK.

ConCloze 2: Claire on Item 4 (Target word: *recreational*)

R: Say it, whatever comes to your mind.

Claire: OK, I should read. [IA]

R: Yeah. Up to you.

Claire: [Line 1, in part, left-hand only] ‘commitments in their lives psst [P] such as family, employment, and [E] other [E] [P] [KWIC position here]’ OK, so the [P] answer could be one of er [Line 1 again immediately, in part] ‘commitments in their lives [P] other [P] [KWIC position here] activities and hobbies that capture their interest.’ OK, so I will [P] I will see ans [P] examples, [Option A] ‘avian’, [Option B] ‘fluffy’, [Option C] ‘mechanical’, [P] er [P] [Option D] ‘recreational’. [P] Can I ask you er the meaning of [P]

R: No.

Claire: Hahaha [laughing]

R: Sorry.

Claire: I don’t know. I don’t know the meaning of number one, A [i.e., Option A ‘avian’] a-vi-an. Yeah, anyway, I think, yeah. If I possibly er [P] say it, the answer is number D [i.e., Option D ‘recreational’] recreational.

R: Why?

Claire: Because er

R: You have read only one line.

Claire: Yeah, the ‘capture’ [from Line 1], the hinting word is er ‘capture their interest’ yeah, so, hmm

R: Uh-huh. Why? What do you mean by ‘capture interest’ [from Line 1]? So what?

Claire: It’s er hmm [P] It’s a something er [P] to make [P] our interest [P] yeah. So, [P] Anyway, I will read more. [Line 1 continued, in phrase] ‘The college players’ [Line 2, in part] ‘frequent [P] [‘use’ not verbalized] If their use of these drugs is [E] often [E] [P] blank [i.e., KWIC position] and associated with leisure acti[vities]’ I think yeah it’s er hinting, or it’s er ‘leisure activities’ [from Line 2]

R: Uh-huh.

Claire: So, [P] [Line 2 again immediately, in word, right before the KWIC position] ‘often [Option D] ‘re [P] creational’’

R: Uh-huh.

Claire: That’s it.

R: O, alright. So, you have read only two lines and can answer. So, the words that help you is [P] ‘capture their interest’ [from Line 1] and ‘leisure activities’ [from Line 2]. Alright. Good.

ConCloze 2: Dakota on Item 4 (Target word: *recreational*)

Dakota: [Option A] ‘avian’, [Option B] ‘fluffy’, [P] hmm [Option C] ‘machin machinical [in fact ‘mechanical’], [Option D] ‘re-create recreational’ The two words I don’t know what [IA ‘they’] mean.

R: Uh, you mean Choice A [Option A ‘avian’] and Choice B [Option B ‘fluffy’]?

Dakota: Yeah,

R: It's alright. OK.

Dakota: OK.

R: So, you [IA 'will'] read the er the choices first.

Dakota: Yeah, I think it will be [P]

R: Faster

Dakota: Yeah, [IA] do they make sense, do they really make sense [IA]

R: OK.

Dakota: [Line 1, in part, left-hand only, with KWIC inserted] 'commit com [IA]'

R: Say it whatever you are thinking.

Dakota: [Line 1 continued] '[IA] lives such as family, employment, and oth and other [P] hmm [KWIC position here] activities' [P] [Option D] 're recreational' [P] [Line 2, in phrase, right before the KWIC position] 'drugs is often [KWIC position here]' [P] I don't know this [E] words [E] [i.e., either Option A 'avian' or Option B 'fluffy' or both]! How can I choose then? [Line 3, in phrase, right before the KWIC position] 'bet[ween] competitive and er [P] [KWIC position here] [IA]' [P] [IA 'I'll guess']. And [Line 4, in part] 'A large number of res respondents also commented commented commented that [KWIC position here] [P] boaters are regarded ['as' not verbalized] fat cats' [P] [Option B] 'fluffy' [P] that's [Option B] 'fluffy' [P] 'fluffy' [Line 5, in phrase, from right before the KWIC position, with KWIC inserted] 'more [Option B] 'fluffy' drugs'. No, psst. [Line 6, in phrase, right before the KWIC position, with KWIC inserted] 'educational, users [P] [IA Option C 'mechanical'?]' [Line 6 again immediately, in phrase, right before the KWIC position, with KWIC inserted] '[IA 'was'] meeting users' educational [P] [Option A] 'a-va-rian' [i.e., 'avian'] a-ver-ian, averian, what is 'a-ve-rian'? I forgot.

R: I cannot say.

Dakota: [Line 6 again immediately, in phrase, right before the KWIC position] 'educational [P] users [KWIC position here]' [Line 6 again immediately, in phrase, right before the KWIC position] 'educational and the, psst, [KWIC position here]' [P] [Line 6 again immediately, in phrase, right before the KWIC position] 'and the [KWIC position here]' [P] [Line 6 again immediately, in phrase, right before the KWIC position] 'educational and the [KWIC position here] [P] needs' [P] [Line 6 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'and the [Option D] 'recreational' needs' [Line 7, in part, from right before the KWIC position] 'sold are for [KWIC position here]' [P] [IA] [P] I really don't know the answer [IA 'anyway'] [P] [Line 7 continued, in phrase, right-hand only] 'sold each year worldwide' [P] so, [Line 7 again immediately, in part] 'to the', ok 'most, [IA] most of [P] bicycles sold [E] are for [KWIC position here] [E] for what [IA 'you'] use' [P] Should be [Option C] 'me-chi-nai [i.e., 'mechanical']' No 'me-chi-nai' [P] me-chi-nai-cl [P] me-chi-nai-cl [P] No? [P] [Line 7 continued] 'million bicycles sold each year worldwide' [Line 7 again immediately, in word, right before the KWIC position, with KWIC inserted] 'for [Option D] 're-creational recreational' Should be this two.

R: What what do you mean 'this two'? C [Option C 'mechanical'] or D [Option D 'recreational']?

Dakota: Yeah, should be this two, but I'm not sure because er because I don't know that two [Options A and B 'avian' and 'fluffy', respectively] er word [P] And also I I don't think er they they, like, to fit in this two sentence.

R: You mean?

Dakota: Two, three [probably Lines 2 and 3]

R: You mean, five to seven?

Dakota: [P] Yeah. Some doesn't like like er this two, put here, doesn't right.

R: Uh.

Dakota: I mean, [P]

R: So, you mean the word in A [i.e., Option A 'avian'] does not go with [P] five [i.e., Line 5], for example?

Dakota: [P] does, I mean, this 'avian' [Option A 'avian'] doesn't go with er [Line 6, in phrase, from right before the KWIC position, with KWIC inserted] 'educational and the [P] [IA Option A 'avian'] needs'

R: So you think it is [P]

Dakota: If you put in the sentence, if they were one [IA 'or'] two er [P] seems not right. So, you can't just put that for [IA 'all'], right?

R: Yeah, it'll be [P] must go to all the lines.

Dakota: Yeah. [P] So the most suitable one for [P] [Line 6 again immediately, in word, right before the KWIC position] 'educational [P] [KWIC position here]' I will go for D [Option D 'recreational'] then

R: OK.

Dakota: Yeah, I do not feel [P] I just feel, I mean [P] I was thinking, If you, if the word doesn't [P] sound right to [E] one [P]

R: Of,

Dakota: Blanket [i.e., KWIC blank]

R: To any one of the line

Dakota: [IA 'Blanket' [i.e., KWIC blank]] Yeah,

R: So, you just reject it.

Dakota: Yeah. I just think, right?

ConCloze 2: Esther on Item 4 (Target word: *recreational*)

Esther: OK, [Option A] 'avian', [Option B] 'fluffy', [Option C] 'mechanical', [Option D] 'recreational'. [P] I'm not really sure about the meaning of this one.

R: [Option A] 'avian'? Uh-huh, it's OK.

Esther: OK. [Line 1, in part, with KWIC inserted] 'commitments in their lives such as family, employment, and other [P] [Option D] 'recreational' activities and hobbies that capture their interest.' [P] Over here, [Option D] 'recreational', it fits very well.

R: Uh-huh.

Esther: OK. [Line 2, in part, with KWIC inserted] 'If their use of these drugs is often [P] [Option D] 'recreational' and associated with leisure activities, then it is likely' OK, here, fits well. Hmm, OK, [Line 3, in part, with KWIC inserted] 'appear to yield different results from comparisons between competitive and [P] [Option D] 'recreational' athletes.' Right? [P] Hmm, [Line 4, in part, with KWIC inserted] 'A large number of respondents also commented that [KWIC position here]' [P] er [P] OK, possibly [Option D] 'recreational' 'boaters are regarded as "fat cats" [P] and are being "picked?" [P] [Line 5, in part, with KWIC inserted] 'defined polydrug use [P] as the use of 3 or [E] [P] [Option D] 'recreational' drugs' and then something like that [chuckle] [Line 5 continued] 'in the past 90 days' OK? [P] [Line 6, in part, with KWIC inserted] 'on how well the library was meeting users' educational and [Option D] 'recreational' needs' [P]

R: Uh, [P] so you seem to [P]

Esther: Yeah, it seems to be. [P] Because they they are not [Line 6 again immediately, in word, right after the KWIC position, with KWIC inserted] "[Option B] 'fluffy' needs"? [An unlikely option in combination with a word from Line 6]. [chuckle]

R: [Chuckle] OK, yeah.

Esther: Uh, [IA 'human'] I don't think "library [IA 'doesn't'] meets [Option C] 'mechanical' needs" [A simple clause formulated from words or elements in Line 6] [chuckle]

R: Hmm, yeah, I don't know.

Esther: Sure!

R: [chuckle] Up to you.

Esther: Because like, 'library' [from Line 6] is [E] for [E] people, [P]

R: Uh-huh.

Esther: And people are not machines, so they should not [Line 6 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'mechanical' needs' [A reasoning formulated from words or elements in Line 6].

R: Alright.

Esther: They should [IA] [IA 'have'] also like [IA 'a technical'] needs [from Line 6] or like 'educational needs' [from Line 6] [IA]. [P] Also [Line 7, with KWIC inserted] 'to the United States where most bicycles sold are [E] for [E] ahem [coughing] [P] [KWIC position here]' Not '[Option C] 'mechanical' use?' [P] 'most of the 105 million bicycles sold each year worldwide' [P] Right [IA] [P] But again this word fits here.

R: OK, [Option D] 'recreational' is the best, OK.

ConCloze 3: Franz on Item 4 (Target word: *recreational*)

Franz: Oh my goshhh [P]

R: Just say whatever you are thinking

Franz: Erm, first of all, [P] I don't know er the exact meaning of the answer [E] A 'avian' [E] [Option A 'avian'], I don't know.

R: Yeah,

Franz: But [P] so, [IA] haha [chuckle] [Option A] 'Avian'? 'Avian'? [P] No, I don't know the meaning. So, [P] I get back to the sentences.

R: Yeah.

Franz: [Line 1, in part] 'commitments in their lives such as family, employment, and other [KWIC position here] tut tut tut' Eh! When you see the word 'activities' [from Line 1], it's [P] it can lead you to the answer D, the Choice D [Option D 'recreational'], [Line 1 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'recreational' activities', it sounds very good.

R: Uh

Franz: But I haven't decided yet.

R: OK [chuckle]

Franz: Erm, [P] [Line 1 again immediately, right after the KWIC position] '[KWIC position here] other activities and hobbies' Ah, you see the word 'hobbies'? That sounds very [Option D] 'recreational' [i.e., pointing to Option D] related haha [laughing] [Line 1 continued, in part, right-hand only] 'that capture their interest. The [E] collage [E] o college, no, [IA] sorry, the college players' [Line 2, in phrase, left-hand only] 'frequent use' [Line 1, in word, right-hand only] 'players' [IA] see the word 'players' [from Line 1] see the word 'hobbies' [from Line 1]. You see, [P] 'activity' [from Line 1] words, right? Erm [Line 2, in part] 'use [IA 'of'] these drugs is [E] often [E] [KWIC position here]' [P] O, this might be adjective, right? Erm [Line 2 continued] '[KWIC position here] and associated with leisure activities'

'leisure' [from Line 2] again, [P] so [chuckle] it can't be [Option B] 'fluffy'. 'Fluffy' doesn't fit here at all.

R: OK.

Franz: You know, you know? nong thep [his native language, approximately meaning Dear Brother Audio-recorder]

R: [chuckle] OK, OK. Go on, please.

Franz: So, [P] psst [E] It can deceive me by the word [Option B] 'fluffy'? [E] [P] Come on!

R: Just say

Franz: Come on! [P] So, I [E] cross [Option B] 'fluffy' out [E] [P]

R: OK, yeah

Franz: See? [P] Erm [P] And you see the word 'athletes' [from Line 3] [P] I can guess from [P] erm [P] the lexicon [P] the lexicon in in this group of sentences [P]

R: Yes.

Franz: And you have 'activities' [from Line 1] you have [P] erm [P] 'players' [from Line 1], right? 'leisure' [from Line 2], 'athlete' [from Line 2], [Line 3, in phrase, right-hand only] '[IA, murmur] superstitious behaviors' Come on! Haha [laughing] [murmur] [Line 4] 'most marinas. A large number of respondents also commented that [P] [KWIC position here] boaters' Eh, 'boaters'! 'Boaters' is a is a kind of [P] 'players' [from Line 1], right? Er it's a kind of 'athlete' [from Line 3]. Erm [P] [Line 4 continued] ' "fat cats" ' Eh, the word "fat cats" must be put here just to confuse [P] [chuckle]

R: OK.

Franz: Just to [P] relate [P] this verb this verb er this word, sorry, to [Option B] 'fluffy'. O, you can't deceive me.

R: OK.

Franz: No! You can't get me! [P] [Line 4 continued] 'picked' [Line 5] 'defined polydrug eh drug use as use of three or [E] more [KWIC position here] drugs [E] in the past 90 years oh 90 days, sorry, methamphetamine, cocaine, [P] crack, MDMA' Hmm? [P] Must be something about sport. [P] So it should not be [Option C] 'mechanical', I think. Erm, [Line 6, in part, left-hand only] 'how well the library was meeting users' educational [E] and [E] [KWIC position here]' [P] [Line 6 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'educational and [Option C] 'mechanical' needs'? [Line 6 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'mechanical' needs' doesn't sound [P] [E] suitable [E]. Doesn't sound [E] proper [E]. Doesn't sound [E] meaningful [E]. [P] [Line 6 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'mechanical' needs', what? Why do people need something mechanical? [P] It can be so if you are crippled. No offence! Right? If you are [P] disabled, let's say, if you are disabled, you must be engaged in something [E] [Option C] 'mechanical' [E] [P]

R: OK.

Franz: But in this context, erm, this group of sentences er [E] deals [E] with erm sports [P] and activities, movement, [P] mobility, I guess. [P] [E] So, now, [E] I have my answer.

R: OK.

Franz: I do choose the answer D 'recreational' [i.e., Option D]

R: OK, anything else that you would like to tell me? About the clues? Or any other elements or components in this item?

Franz: Erm [P] as far as I have [P] observed [P], I can see that the fourth [P] question [P] erm is easier [P] than [P] the all previous questions.

R: Really?

Franz: Because I can group [P] I can group erm lexicon [P] that I [E] found [E] here. [P] erm [IA 'I can use'] past simple, [Self-monitoring?] [IA]

R: Just say it.

Franz: [chuckle] I define it, right? [IA 'can use present simple as well']

R: I don't know. Go on, go on.

Franz: [Chuckle] OK, [P] As I told you, erm you have 'activities' [from Line 1], you have 'hobbies' [from Line 1], you have 'leisure' [from Line 2], you have 'activities' [from Line 2], [P] and you have 'athletes' [from Line 3] and you have [P] 'drugs' [from Line 5] [chuckle]

R: Uh-huh.

Franz: And [P] people who do sports tend to erm [P] get involved in drugs [chuckle] in one way or another. That is so funny. Erm, that's all. So, I chose [Option D] 'recreational'.

R: OK, thank you.

[From the next file]

Franz: I told you before that I [E] cross [E] the word I don't know [E] out [E].

R: Yeah.

Franz: [Option A] 'avian' [P] I don't need it. [P]

R: OK.

Franz: That's my strategy.

R: OK. [chuckle]

ConCloze 3: Gill on Item 4 (Target word: *recreational*)

R: Only five, so this is the fourth [item].

Gill: O, OK, OK, OK.

R: Yeah, yeah, yeah.

Gill: Er [Line 1, in word, left-hand only] 'commitments' here the words are [P] [Option A] 'av avian', [Option B] 'fluffy', [Option C] 'mechanical', [Option D] 'recreational' haha [laughing] [Line 1 again immediately, in part, left-hand only, with KWIC inserted] 'commitments in their lives such as family, employer [in fact 'employment'], and other [P] [IA] [IA Option C 'mechanical'] [Option D] 'recreational' OK here. [Line 1 again immediately, in part, right after the KWIC position, with KWIC inserted] '[Option D] 'recreational' activities and hobbies ['that' not verbalized] capture their interest' OK. So this er 'activities', 'hobbies', and 'capture their interest' [all from Line 1] have many to [IA 'go in'] for [Option D] 'recreational'

R: Wow, from the first line.

Gill: Yeah, [P] er [Line 2, in part, left-hand only] 'frequent use [IA 'of] [P] If their use of these drugs is often [KWIC position here]' yeah, I I do know that [P] [Line 2 again immediately, in phrase, left-hand only] 'If their use of drugs' is either [IA 'as'] medicine or as recreational [the word from Option D 'recreational'].

R: Hmm.

Gill: It's [IA 'a trend for African world'?], so I'm going for [Option D] 'recreational'.

R: Uh-huh.

Gill: The drugs is often recreational [the word from Option D 'recreational'], so it suits with 'leisure' [from Line 2], OK.

R: Uh.

Gill: [Line 3, in part] 'It ep appear to yield different results from comparisons [P] between competitive [P] and [P] er [KWIC position here]. Furthermore' if yeah [P] er if if [P] If something is not competitive [word from Line 3] it may it is 'leisure' [from Line 2] or [Option D] 'recreational'.

R: Hmm.

Gill: So, I will go for, [Line 3, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'recreational' athletes'. Some people are er a I I'm 'athlete' [the same lemma as word from Line 3] can either be for the sake of co competition [the same lemma as word from Line 3] or for the sake of [P] [Option D] 'recreational'.

R: [Option D] 'recreational'

Gill: Yeah

R: Uh-huh.

Gill: [Line 4] 'most marinas. [P] A large number of respondents also commented that [KWIC position here] [P] boaters are regarded as "fat cats" and are being "picked' [Line 4 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'commented that [P] [IA Option D 'recreational'] boaters [IA] fat cats [IA]' Psst [P]

R: Keep saying.

Gill: [IA] [Line 4 again immediately, in word, right before the KWIC position] 'commented [KWIC position here] [IA]' [Line 4 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'respondents also commented that [P] [Option D] 'recreational' boaters [IA] [P] fat cats and are being picked' [P] OK, I'll I'll go forward and come back.

R: Yeah.

Gill: [Line 5, with KWIC inserted] 'defined polydrug use [P] as the use of 3 or [E] more [E] [P] yeah [Option D] 'recreational' drugs in the past' yeah [P]

R: Uh-huh.

Gill: Erm [Line 5 continued] 'meth-phe meth-amphetamine, cocaine, crack, [P] and MD[MA]' OK.

R: Uh.

Gill: Er [Note that Line 6 is not verbalized here, potentially unintentionally.] [Line 7, in part, with KWIC inserted] 'to the United States [P] where most bicycles are sold for [P] of course, [Option D] 'recreational' use'

R: Uh-huh.

Gill: So, the the er come back to four [i.e., Line 4]. Four is [IA 'presenting me'] [IA] with a bit of [P] thinking [IA] [sigh]

R: Yeah.

Gill: I need to think. [P] [Line 4, in part, left-hand only] 'A large number of respondents also commented that [KWIC position here]' [P] Well, OK, I'll I'll say that 'boating' [word derived from 'boaters' in Line 4] [P] is a sport.

R: Uh-huh.

Gill: I'll say that 'boating' [word derived from 'boaters' in Line 4] is a sport, and people would go out er go boating, [P] so [P] maybe, they're [P] it's it could be [Option D] 'recreational'.

R: OK.

Gill: [Line 4 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'recreational' boaters'. So, I will go for [Option D] 'recreational'.

R: OK. Perfect.

ConCloze 3: Halle on Item 4 (Target word: *recreational*)

Halle: OK, so, [sigh] [P] first line, [Line 1] ‘commitments in their lives such as family, employment, and [E] other [E] [P] blank [i.e., KWIC position] activities and hobbies that capture their interest. The college players’ [P] and the words are [P] [Option A] ‘avian’? I guess. I’d I think it is for some [IA ‘airways’] for this word, so that is a bit problematic, [P] [Option B] ‘fluffy’, [Option C] ‘mechanical’, or [Option D] ‘recreational’. [P] So, [Line 1 again immediately, in part, left-hand only] ‘commitments in their lives such as family, employment, and [E] other [E] [P] [KWIC position here]’, I think, [Line 1 again immediately, in part, right-hand only, with KWIC inserted] ‘[Option D] ‘recreational’ activities and hobbies that capture their interest.’ [P] It’s a bit problematic because I don’t know what [Option A] ‘avian’ thinks er means, but [P] I’m pretty sure for this one. [Option D] ‘recreational’ fits. I’m gonna go with that one, for now. [P] [IA] with the second one. So, second one [Line 2] ‘frequent use. If their use of these drugs is often [P] blank [i.e., KWIC position] and associated with leisure activities, then it is likely that’. Definitely, [Option D] ‘recreational’ because it’s such er use erm same technique, ‘[Option D] ‘recreational’ use of drugs’ [from Line 2, restructured with the KWIC selected], obviously goes there. And the third line [Line 3] ‘appear to yield different results from comparisons between competitive [E] and [E] [P] blank [i.e., KWIC position] athletes. Furthermore, an examination of specific categories of [E] supperstit [E] supper [IA] [P] superstitious behaviors’ [P], so, again, [P] [Option D] ‘recreational’, [Line 3 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] ‘competitive and [Option D] ‘recreational’ athletes’, fits really well. [P] psst, fourth line, [Line 4] ‘most marinas. Hashtag, a large number of respondents also commented [E] that [E] blank [i.e., KWIC position] boaters are regarded as “fat cats” and are being “picked” I think again, [Option D] ‘recreational’ works. [Line 4 again immediately, in word, right after the KWIC position, with KWIC inserted] ‘[Option D] ‘recreational’ boaters’, especially because it is [Line 4 again immediately, in phrase, right-hand only] ‘are regarded as “fat cats” ’ [Relevant?], so [P] that kind of being linked with [Option D] ‘recreational’ in a way [chuckle] [Inconceivable logic!].

R: Yeah.

Halle: Then fifth line, [Line 5] ‘defined polydrug use as the use of 3 or [E] more [E] [P] blank [i.e., KWIC position] drugs in the past 90 days meth-amphetamine, cocaine, crack, MDMA’ Again, [Line 5 again immediately, in word, right after the KWIC position, with KWIC inserted] ‘[Option D] ‘recreational’ drugs’. No doubt about this [E] one [E]. [P] Six. [Line 6] ‘on how well the library was meeting users’ [E] educational [E] and blank [i.e., KWIC position] needs so improvements could be made. They agreed to complete’. So again, [Line 6 again immediately, in word, right after the KWIC position, with KWIC inserted] ‘[Option D] ‘recreational’ [E] needs [E]’, works really well. [P] Then finally, [Line 7] ‘to the United States, where most bicycles sold are [E] for [E] blank [i.e., KWIC position] use most of the one hun [P] dred million bicycles sold each year worldwide.’ Again, [Line 7 again immediately, in word, right after the KWIC position, with KWIC inserted] ‘[Option D] ‘recreational’ use’, coz it is quite [IA] familiar [P] expressioning, for example, [IA ‘around’] ‘bikes’ [‘bicycles’ in Line 7]’. So, I think it’s work.

R: OK.

ConCloze 3: Igor on Item 4 (Target word: *recreational*)

R: [IA] Comes to your mind.

Igor: [Line 1, in part] ‘commitments in their lives such as family, employment and other [P] [KWIC position here] activity’ [Line 1 again immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘and other [P] [IA [Option D] ‘recreational’], [P] [Option A] ‘avian’, [Option B] ‘fluf[fy]’, this I don’t know much about vocab, so

R: Yeah.

Igor: [Line 1 again immediately, in phrase, from right before the KWIC position] ‘and other [KWIC position here] [E] activities [E]’ should be adjective. [P] So, I am looking for [P] adjective. [P] [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] ‘[Option C] ‘mechanical’ activity [in fact ‘activities’] and hobby’. [P] ‘[Option D] ‘recreational’ OK, I choose er D [Option D ‘recreational’], I think, the first one [i.e., Option D ‘recreational’, based on the engagement with Line 1]. [Line 2, in part, left-hand only] ‘frequent [P] use. If their use er of these drugs [P] is [E] often [E] [KWIC position here]’ [P] [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] ‘is often [P] [Option D] ‘recreational’ and associated’ [P] Yeah, [P]

R: So you have read just [P] one line and then you decided?

Igor: No, no, no, I cannot. It’s er It’s er not correct anymore. [P] [Line 2 again immediately, in part, left-hand only] ‘If their use of these [E] drugs [E] is often [KWIC position here]’ [P] [Line 2 again immediately, in phrase, right before the KWIC position] ‘is often [KWIC position here]’, should be [E] verb [E], here. [P] Should be pas passive voice. [P] [Line 2 again immediately, in phrase, from right before the KWIC position] ‘[IA ‘often’] [KWIC position here] [P] and associate [IA]’ OK, let me see [IA ‘that’] all. [P] [Note that Line 3 seems to be skipped here, or else it must be read very swiftly in silence and thus not verbalized.] [Line 4, in part, with KWIC inserted] ‘most mari most mari[nas] [IA]’

R: Keep saying.

Igor: [Line 4 continued] ‘comment[ed] that hmm [P] [Option D] ‘recreational’ [P] boaters are regarded [P] [IA]’

R: Keep saying.

Igor: [Chuckle] [P] Yeah, this one, OK, I choose, [P] [Line 5, in phrase, from right before the KWIC position, with KWIC inserted] ‘more [P] [Option D] ‘recreational’ drug,’ [P] [Option C] ‘mechanical’, [IA] I think, OK, should be D [Option D ‘recreational’], I guess. [chuckle]

R: Why?

Igor: This one, I just er see the structure of the sentence, and I see some [P] some vocab missing

R: For example?

Igor: Yeah.

R: What do you mean, ‘structure’

Igor: Yeah.

R: Of the sentence?

Igor: Yeah, this one if I don’t know about the the meaning of the vocab, so I will s see the structure. Like [P] sometime[s] verb, sometime[s] adjective missing. So, this one you need adjective.

R: Uh-huh.

Igor: So, they have two word[s]: it [P] its look sound good when you [P] put it in the question

R: Which line exactly?

Igor: Erm

R: And what are the key words

Igor: [P] hmm

R: That help you?

Igor: [P] It, all of them it can apply [P] [IA ‘at’] [P] I think the first one, yeah, the first one.

R: The first one?

Igor: Yeah, [IA ‘choose’] [P] [IA ‘or’] [P]

R: Why? They’re there are twenty words [P] in Line 1, so which words are significant?

Igor: [Chuckle] No, I just I just [P] sense. I think it's er

R: I mean, to you, coming to the conclusion

Igor: [Option D] 'recreational'

R: That [E] D [E] should be the answer.

Igor: [IA] 'activity' [from Line 1], I think. About 'activity' [from Line 1] and 'hobby' [from Line 1].
Something you need to [re]creation.

R: What do you mean?

Igor: [P] Hmm, yeah, this this one, key word for me.

R: Why? Why do you think it's key word for D [Option D 'recreational']?

Igor: Hmm.

R: Could you please explain?

Igor: Erm when you apply [P] [Option D] 'recreation[al]' with 'activity' [from Line 1], [P] it's look make
sense for the meaning for me.

R: Make sense?

Igor: Yeah.

R: What do you mean, 'make sense'?

Igor: Sometime you [P] push ['put'?] someone ['something' or 'some word'?], but it's not familiar or, [P]
yeah, but this one, it's look familiar for me, yeah.

R: Alright.

Igor: I don't know. I I

R: No worries.

Igor: I just guess it. It's the way to guess.

ConCloze 4: James on Item 4 (Target word: *recreational*)

R: Whatever you are thinking.

James: [Line 1, in part, left-hand only] 'commitments in their lives, such as family, employment, and
[KWIC position here]' [P]

R: Yeah,

James: And

R: Keep saying.

James: Yeah, 'act' [the following word in Line 1, 'activities'] [Line 1 again immediately, in part, from right
before the KWIC position] 'other blank [i.e., KWIC position] activities and hobbies that capture their
interest. The college players' [P] [Line 2] 'frequent use. If their use of these drugs is often blank [i.e.,
KWIC position] and associated with leisure activities, then it's likely that' [P] [sigh] [Line 3] 'appear
to yield different results from comparisons between competitive n and blank [i.e., KWIC position]
athletes. [P] Furthermore, an examination of specific categories ['of' not verbalized] superstitious
behaviors' OK. [P] Er [Option D] 'recreation[al]' the [IA] [P] [Option A] 'avian', [Option B] [E]
'fluffy', [Option C] 'mechanical' and [Option D] 'recreational'. [P] [Line 1, in part, left-hand only]
'com[mit]ments in their lives such as family, employment, and other [KWIC position here]' [P] [Line
1 again immediately, in part, left-hand only] 'com[mit]ments in their lives such as family, employment,
and other [KWIC position here]' [P] Hmm

R: Keep saying.

James: Yeah. Er I'm just er [chuckle]

R: Yeah, whatever you are reading.

James: [Line 1 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'and other [Option C] 'mechanical' activities and hobbies' It's OK. [Line 2, in part, with KWIC inserted] 'freq [P] if their use ['of' not verbalized] these drugs ['is' not verbalized] often [P] [Option C] 'mechanical'? And the' No, [IA 'it's not very'] er, [P] Actually, I don't know the meaning of these two words haha [laughing]

R: Uh, you mean Choice A [Option A 'avian']? Uh-huh, uh-huh

James: Yeah, A [Option A 'avian'] and B [Option B 'fluffy'], yeah.

R: Yeah.

James: It's er [IA] idea. Maybe, the [Option D] 'recreational' [P] and [Option C] 'mechanical' might not be the answer because er [P] here, [P] the, Line 2, [IA 'there're'] [Line 2 again immediately, in phrase, right after the KWIC position] '[KWIC position here] and associated with leisure activities' [IA] It's kind of same same [P] meaning. [P] So if I choose [Option D] 'recreation[al]' here, [P] and er those two are going to be the same meaning. So, [P] should be something different from the [Option D] 'recreational' because er [P] same answer should appear the the all seven lines. So, [P] maybe, one of these, I'm I'm not sure about the

R: Uh-huh.

James: The meaning of this [IA]

R: Uh-huh.

James: So, er, choose, can I choose?

R: Up to you! So, if if, so you would like to guess [with the meaning of one of the options unknown]?

James: Yeah.

R: Yeah.

James: So, [Option B] 'fluffy' [P]

R: So, how will you guess? Which one?

James: Er [P] [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] 'employment, the [in fact 'and other'] [Option A] 'avian' activities' er it's even difficult to guess, because er I'm totally, I don't know what's the meaning of this two.

R: Uh-huh, uh-huh.

James: So, so, it's just er like er I'm [IA]

R: Yeah, so, OK, you like that word more.

James: Yeah, yeah, yeah.

R: Yeah, OK, no worry.

ConCloze 4: Klavier on Item 4 (Target word: *recreational*)

R: Whatever you are thinking.

Klavier: Er, [Line 1] 'commitments in their lives such as family, employment, and other [P] [KWIC position here] activities and hobbies that capture their interest.' Erm, I'm not sure what [Option A] 'avian' means. [P]

R: Er, no worries.

Klavier: Erm, [chuckle] OK. [Line 1 continued] ‘The college players’ erm [Line 2] ‘frequent use. If their use of these drugs is often [KWIC position here] [P] and associated with leisure activities, then it is likely that’ [P] [Line 3] ‘appear to yield different results from comparisons between competitive [E] and [E] [KWIC position here] [P] athletes. Furthermore, an examination of specific categories of [P] superstitious behaviors’ [P] [Line 4] ‘most marinas. A large number of respondents also commented that [KWIC position here] [P] boaters are regarded as “fat cats”’ [Recording interrupted]

Klavier: Erm [P] Because because of ‘drugs’ [from Line 2], erm, I heard that, I heard drugs and [Option D] ‘recreational’ in the same erm context.

R: Uh-huh.

Klavier: Erm before.

R: Ah.

Klavier: And then, more [P] er [IA ‘frequencies’??] of erm of ‘drugs’ [from Line 2], so I think it’s [Option D] ‘recreational’.

R: Uh-huh.

Klavier: Well, I’m not quite sure what [Option A] ‘avian’ means, but I I don’t think it’s that. I think it’s [Option D] ‘recreational’. [P] Erm [P] yeah.

R: Yeah.

[...]

Klavier: Erm [P] Because because of ‘drugs’ [from Line 2], erm, I heard that, I heard drugs and [Option D] ‘recreational’ in the same erm context.

R: Uh-huh.

Klavier: Erm before.

R: Ah.

Klavier: And then, more [P] er [IA ‘frequencies’??] of erm of ‘drugs’ [from Line 2], so I think it’s [Option D] ‘recreational’.

R: Uh-huh.

Klavier: well, I’m not quite sure what [Option A] ‘avian’ means, but I I don’t think it’s that. I think it’s [Option D] ‘recreational’. [P] Erm [P] yeah.

R: Yeah.

ConCloze 4: Lulu on Item 4 (Target word: *recreational*)

Lulu: Hmm [Option A] ‘avian’, [Option B] ‘fluffy’, [Option C] ‘mechanical’, [Option D] ‘recreational’. Hmm, interesting combination haha [chuckle] [P] [Line 1] ‘commitments in their lives such as family, employment, [P] and [E] other [E] [KWIC position here] [P] activities and hobbies that capture their interest’ [P] [Option D] ‘recreational’, right? [P] [Line 1 again immediately, in part, left-hand only] ‘commitments in their lives such as family, employment, and [KWIC position here] [IA ‘no’] [IA]’ [P] It’s ‘commitments’, ‘family’, ‘employ[ment]’ [P] [IA ‘cap[ture]’] [all from Line 1] [Option D] ‘recreational’. [Line 1 again immediately, in part, right-hand only, with KWIC inserted] ‘[Option D] ‘recreational’ activities and hobbies [P] that capture their interest.’ [P] [IA ‘psst, interesting’] [Option D] ‘recreational’. [Option C] ‘mechanical’ doesn’t make sense. [Option B] ‘fluffy’? I don’t think it goes with ‘activities’ [from Line 1]. [Option A] ‘avian’ [P] Not sure [IA ‘what’] I know what it means. [chuckle] [P] [Line 2, in phrase, left-hand only] ‘frequently [in fact ‘frequent’] use. If their’ I just see if erm psst [Option D] ‘recreational’ fits [IA] with [IA] sentences, [IA] ‘drugs’ [from Line 2] and I will go for that. [IA] [P] [Line 2 again immediately, in part, with KWIC inserted] ‘If their use of these drugs is often [Option D] ‘recreational’ and associated with [E] leisure [E] activities’ Yeah, [Line 2 again

immediately, in part, right-hand only, with KWIC inserted] [Option D] ‘recreational’ is associated with leisure activities’ [IA] Yes, [IA ‘I’m going for’] this one.

R: Yeah.

ConCloze 2: Aaron on Item 5 (Target word: *livestock*)

R: [IA] To your mind, whatever you are thinking. We want to know what you are thinking.

Aaron: Yeah, I think I’m lucky because this kind of question is not in my IELTS [P] test. [P] Erm, [Line 1, in part, left-hand only] ‘nineteen sixty-nine, nineteen [IA]’ [E] [Option A] ‘breeding’, [Option B] ‘creature’, [Option C] ‘livestock’, [Option D] ‘stray’ [E] [P] Hmm, [P] [Note that Line 2 does not appear to be verbalized here—probably skipped altogether] [Line 3, in phrase, right before the KWIC position] [E] ‘their stolen [E] [KWIC position here]’ [P] [Line 4, in word, right before the KWIC position] [E] ‘slaughtering [E] [KWIC position here]’ [P] [Line 4 again immediately, in word, right before the KWIC position] [E] ‘slaughtering [E] [KWIC position here]’ [Line 4 again immediately, in word, right before the KWIC position] [E] ‘slaughtering [E] [KWIC position here]’ [P] [Line 5, in phrase, right before the KWIC position] [E] ‘restocking [‘of’ not verbalized] [KWIC position here]’ [Line 6, in phrase, right before the KWIC position] [E] ‘increase in [KWIC position here]’ [P] er [Line 7, in word, right before the KWIC position] ‘the [KWIC position here]’ [P]

R: Just say it. Keep saying.

Aaron: Yeah, yeah. [P] [Note that two concordance lines appear not to be verbalized or else skipped altogether here.] [Line 10, in phrase, right before the KWIC position] ‘escape with some of his [KWIC position here]’ [P] No. I don’t [P] I don’t know what what is [Option C] ‘livestock’ [Also the key]. [P] It it [P] [Option C] ‘livestock’, what is ‘livestock’? [Does this mean that he has tried other choices but they are unlikely?] [P]

R: Just say whatever you are thinking.

Aaron: Yeah, yeah, yeah. [P] I [P] [Option A] ‘breeding’ for me is is not inte is not hmm [P] it’s not noun. But the other choices erm I think they they are [P] hmm they are nouns. But the first choice is not noun. [P]

R: Uh-huh.

Aaron: So I [P] I tend to choose [P] er among the last three words [i.e., options provided: Option B ‘creature’, Option C ‘livestock’, and Option D ‘stray’].

R: You mean, from B to D dog, right?

Aaron: Hmm. [P]

R: So, which word

Aaron: Which word

R: Helps you? [P]

Aaron: [Line 4, in word, right before the KWIC position] ‘slaughtering [KWIC position here]’ [P], I think. When I see this kind of words, [P] it should be [IA ‘at least’] as something like [Option B] ‘creature’.

R: Uh-huh.

Aaron: But I am not really familiar with the use of [Option C] ‘livestock’.

R: Uh-huh.

Aaron: And if I choose [Option B] ‘creature’ and put it in the seventh sentence, [Line 7, in phrase, from right before the KWIC position, with KWIC inserted] ‘the [Option B] ‘creature’ producer’ is not very appropriate [P] to for me.

R: Uh-huh.

Aaron: Er [P] so, [P] [Option D] ‘stray’ psst, [P] Oh, [Option D] ‘stray’ is also strange. [P] I seldom use this words in in my research. Hmm.

R: Uh-huh.

Aaron: So, my feeling, psst, [Line 3, in phrase, right before the KWIC position] ‘their stolen psst [KWIC position here]’ [P] [Line 3 again immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘their stolen [Option B] ‘creature’ [P] psst [IA ‘raids’] [IA]’ [Line 3 again immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘their stolen [Option C] ‘livestock’ also OK. [P] Hmm, [P] [Line 9, in phrase, right before the KWIC position, with KWIC inserted] ‘co continued to raise [Option B] ‘creature’ is not appropriate. So, [P] I think [P]

R: [Option C] ‘livestock’?

Aaron: [P] Yeah I don’t know. This this questions [‘concordance lines’?] are very hard for me. Yeah.

R: Alright.

Aaron: Yeah. All the questions. Be besides, er, in fact, for the the first two questions, all all the all this kind of questions are hard for me to answer.

R: Uh-huh, OK, thank you so much.

Aaron: Finish?

R: Yeah.

ConCloze 2: Björn on Item 5 (Target word: *livestock*)

Björn: Then [P] Again, [P] this, the last one for me. [chuckle]

R: Just say whatever comes to your mind.

Björn: [Line 10, in part, left-hand only] ‘injure [IA ‘his family, and then escape’] some of his [P] [KWIC position here]’ [NB: By left-hand appearance, this Line 10 appears to be the shortest one.] er [Option B] ‘creature’, [Option C] ‘livestock’ [IA ‘and then we turn for’] [IA ‘have some kind of the’] [IA] [Line 1, in part, left-hand only, with KWIC inserted] ‘droughts [IA ‘to’] have’ [IA] [in fact, the line says ‘However, the droughts in 1969–73 and 1984 have considerably diminished’] [P] [Option A] ‘breeding? breeding’? [IA ‘so, find the words’] [Line 10, in word, left-hand only] ‘injure’ in the [IA ‘sentence text’]. [P] [IA] are ahem [coughing] [Line 8] ‘the local lumber [IA ‘yard’], they could contract [P] to ship their [P] [KWIC position here] to South St. Paul’ [P] yeah, [P] [Option A] ‘breeding’? and [Line 8 continued, almost IA] ‘they could deal with local grain’ [P] [Line 8 again immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘to ship their [Option B] ‘creature’ [IA ‘[Line 10, in phrase, right before the KWIC position, with KWIC inserted] ‘escape with some of his [Option B] ‘creature’]’ [P] [IA] [Line 7, in part, left-hand only, with KWIC inserted] ‘Even though [IA ‘that’] animal is likely to recover’ er [P] for this er [Option D] ‘stra stray’? [P] [Line 6, in word, right after the KWIC position, with KWIC inserted] ‘[Option D] ‘stray’ numbers’ er No, [Option D] ‘stray’. [P] Er [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] ‘have considerable dis dimi diminished [P] [Option D] ‘stray’? [P] and pastures’ [IA ‘same, same, same?’] [NB: the parallel structure created by the presence of the coordinating conjunction ‘and’] [IA] [P] [Line 2, in phrase, from right before the KWIC position] ‘against this [IA] [KWIC position here] disease’ [IA] [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] ‘this [Option D] ‘stray’ disease’? [P] [IA] [P] Hmm, [P] OK, I think [IA] put an [P] adjective? Here? [Line 1, in word, right before the KWIC position] ‘diminished [KWIC position here]’ No, a noun. A noun. [Line 1 again immediately, in word, right after the KWIC position] ‘[KWIC position here and ‘and’] pastures’ [P] So, from this [P] [Line 2, in word, right after the KWIC position, with KWIC inserted] ‘[Option B] ‘creature’ disease’ [P] Erm, yeah, [P] [Line 3, in phrase, from right before the KWIC position, with KWIC inserted] ‘their stolen [Option B] ‘creature’ from’ [IA] This the most make sense one. [P] for question three? [i.e., Line 3 seems to give the best clues.] [P] and [P] one? [i.e., Line 1 also functions as

the most hinting of the probable answer.] But then look all, I haven't look for all sentences. Still, I choose B [Option B 'creature'] for the moment.

R: OK. Moment, moment. Just

Björn: Yes?

R: Er, when you do this kind of question,

Björn: OK,

R: How do you start? Do you read from the first line?

Björn: Erm,

R: Do you always read from the first line?

Björn: Actually not. Erm but usually I will read from the first line. Erm

R: Uh-huh.

Björn: It's [P] normally erm ask for people to read from start, but er sometimes if I feel nervous or, [P] er, yeah, if I feel nervous, I will try to look, [P] try to change my ways, maybe I look from down, from bottom to the top. Or maybe er only a part of the sentence. So erm, yeah, it's not usually look from the top but erm

R: Yes.

Björn: It's depends on [P] the emotion haha [chuckle], yes

R: Er from Question 1 to Question 5, do you read all the lines to get to the answer or do you just read a few lines [P]

Björn: Erm

R: And then you get, and then you come to the answer already?

Björn: Actually I read all the lines,

R: Yes.

Björn: Because I want to have erm a fully understanding of the question.

R: Yeah.

Björn: I try to

R: To get as much as information

Björn: Yes, but erm, yeah, I will read all the lines.

R: And, [P] as for this question five, which word, [P] which word helps [P] helps you to get to the answer?

Björn: Er

R: All the words? Or just some specific words?

Björn: Erm, hmm, a few words. Actually I look for the erm

R: [from what has been observed] [Line 1, in word, right after the KWIC position] '[KWIC position here] pastures'?

Björn: Yes, [Line 1, in word, right after the KWIC position] '[KWIC position here] pastures', and [Line 3, in phrase, right before the KWIC position] 'to exchange their stolen [KWIC position here]'. [P] Especially Sentence one and Sentence three [i.e., Lines 1 and 3] give me the hint of words of what are

R: How?

Björn: Simi similar words [IA] 'pastures' is er [P] nec. [P] Actually I looking at the at the nature of words.

R: Uh-huh. But they [i.e., the options] are all nouns. I can guarantee.

Björn: Yes, they are all nouns, so I list, choose the most similar one.

R: Similar? What do you mean? They are [P], all the ten all the ten sentences here want only one word, and they are all the same word.

Björn: Yes, yes, yes.

R: So?

Björn: Er, hmm.

R: So, in the first line you look at what word? [Line 1, in word, right after the KWIC position] '[KWIC position here] pastures' and [P] any any other words?

Björn: [Line 3, in phrase, right before the KWIC position] 'their stolen something [i.e., KWIC position]' [Line 3 again immediately, in phrase, right before the KWIC position] 'their stolen [IA] [KWIC position here]'

R: [Line 3, in phrase, right before the KWIC position] 'their stolen dot dot dot [i.e., KWIC position]'

Björn: [Line 3 continued, in phrase, right after the KWIC position] '[KWIC position here] from raids in Souvenirs [in fact 'Sonora']' [P] [IA 'third'] sentence [i.e., Line 3]

R: Uh-huh.

Björn: And also [Line 5, in phrase, from right before the KWIC position] 'restocking of something [i.e., KWIC position] in Tes [P] [Te]so'

R: Yeah.

Björn: OK, I think [P] it can be [Option C] 'livestock' or [Option B] 'creature cre cre crea'

R: [Option B] 'creature', uh-huh.

Björn: [Option B] 'creature'

R: And what what about sentence ten or nine [i.e., Lines 9 and 10], do they help you?

Björn: I [P] I don't think so, because [P] here the word 'injure' [from Line 10]

R: 'Injure' [from Line 10] in sentence 10 and

Björn: Make me [chuckle] think erm [IA 'answering'] [Option A] 'breeding' because people [IA] linking the erm words [Option A] 'breeding' 'injure'[from Line 10], right? [IA] [NB: Probably confusion of [Option A] 'breeding' with 'bleeding']

R: Uh-huh.

Björn: And er sorry, and I look for the sentence and I choose the most one. It sound the like er to easy to understand it.

R: Uh-huh.

Björn: That is [IA] the most suitable one.

R: What about sentence four [i.e., Line 4]?

Björn: Yes?

R: Does it help you?

Björn: [Line 4, with KWIC inserted] 'available in the state food market, and anyone caught sta sta staughtering [in fact 'slaughtering'] and [P] [Option B] 'creature'? illegally may be impr imprisoned for up to four years' [P] Yes, erm it's [P] [IA 'help me the'] also decide, the nature of word help me to decide.

R: Do you know [P] this word?

ConCloze 2: Claire on Item 5 (Target word: *livestock*)

R: The last one!

Claire: The last one? O, haha [laughing]

R: OK, so say it loudly, whatever you are thinking

Claire: Yeah, I'll I will read from the first sentence. [Line 1] 'However, the [P] droughts in 19 [IA] [the year not verbalized] have er considera[bly] diminished diminished [IA] something [i.e., KWIC position] and er [P] pastures [P] and pastures thereby weakening some er parental elderly and chiefly leverage' er [P] I think I'll see the examples [the options??] er [Option A] 'breeding', [Option B] 'creature', [Option C] 'livestock', [Option D] 'stray' hmm psst [P] I think [Option A] 'breeding', [Option B] 'creature', [Option C] 'livestock' psst. Can't [IA] I will read more. [Line 2, in part] '['island' not verbalized] laboratories [P] where they search for new weapons against this [P] something [i.e., KWIC position] disease [P] that threatens [P] no consumer's health [P] but every consumer ['s pocketbook' not verbalized]' I think it's er [Option C] 'livestock'

R: Why?

Claire: I think it's er it's er the answer should be related er [IA 'to'] the word 'pastures' [from Line 1]

R: Uh-huh.

Claire: And then er [P] it's er 'no consumer's health' [from Line 2] [P]

R: So what? What do you mean, 'consumer's health' [from Line 2]?

Claire: So, [P] something something related er [P] to 'consumer's health' [from Line 2] er [P] with the [IA 'end'] [P] 'pastures' [from Line 1] yeah?

R: Uh-huh.

Claire: So, I think it's [Option C] 'livestock'.

R: OK, any other thing you would like to say?

Claire: OK, I will read more. [Line 3, in part] 'trade allowed [P] ['Chiricahua and' not verbalized] other ['Apache' not verbalized] exchange their stolen something [i.e., KWIC position] from raids ['in Sonora and Chihuahua for the' not verbalized] er food ['guns' not verbalized]' [P] er [Line 4, in phrase, right before the KWIC position] 'food markets and ['anyone' not verbalized] caught [P] slaughtering [KWIC position here]' [P] I think it's 'slaughtering' [from Line 4]

R: What do you mean? So what?

Claire: I think it's er

R: What it has to do with this one?

Claire: 'slaughtering' [from Line 4] I think it's mean killing something?

R: Yeah?

Claire: And then it's er it's related to food

R: 'Food'? 'Food market'? [from Line 4]

Claire: I think it's er [Option C] 'livestock'.

R: Alright. That's all? OK.

ConCloze 2: Dakota on Item 5 (Target word: *livestock*)

Dakota: So, [Option A] 'breeding', [Option B] 'creature', [P] [Option C] 'livestock', [P] [Option D] 'stray'.
OK. [P] [sigh] [Line 1, in part] 'However, ['the droughts in 1969-73' not verbalized] so have

considerably dim diminish diminished [KWIC position here] and the pastures' [Line 1 again immediately, in word, right before the KWIC position] 'diminished [P] [KWIC position here]' OK. [Line 2, in part, left-hand only, with KWIC inserted] 'island laboratories where they search for new weapons against [P] this [Option B] 'creature'' [Line 2 again immediately, in phrase, right before the KWIC position] 'against this [P] [KWIC position here]' [Line 2 again immediately, in word, right before the KWIC position, with KWIC inserted] 'this [Option B] 'creature'' [P] [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'against the this [Option B] 'creature' disaster? [in fact 'disease']' [P] and er [Line 3, in phrase, right before the KWIC position] 'their stolen [P] [KWIC position here]' [Line 3 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'their stolen [P] [Option B] 'creature' from'? [P] [Line 3 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option A] 'breeding' from'? [Line 3 again immediately, in word, right after the KWIC position, with KWIC inserted] '[IA 'stolen their'], doesn't. [Option C] 'livestock'' [P] [Line 4, in phrase, from right before the KWIC position, with KWIC inserted] 'slaughtering [Option C] 'livestock' illegally' Yeah, I think it's the er

R: [Option C] 'livestock'? [P] Why? Again.

Dakota: Seem like er like er two and three. It's about the something. It's about something. It's about the [P] animal. [P]

R: Uh-huh.

Dakota: Seems like about animal.

R: Uh. [P] Which words tell you that it's about animal?

Dakota: This this one, [Line 4, in part, from right before the KWIC position] 'slaughtering [KWIC position here]'. [P]

R: Uh-huh.

Dakota: [Line 4 continued] 'illegally may be [P] [IA]' [P] Maybe this two?

R: 'Creature' [Option B] or 'livestock' [Option C]?

Dakota: [Option B] 'creature' and [Option C] 'livestock'. And the [Line 5, in phrase, from right before the KWIC position, with KWIC inserted] 'restock[ing] of [P] [Option C] 'livestock' in [IA] in Taiso? Teso'

R: I think it's the the city.

Dakota: O, [P] so, [Line 6, in part, with KWIC inserted] 'which you [in fact there is no 'you' here] can lead it [P] large increase in [P] in [Option C] 'livestock' numbers' [Line 7, in phrase, left-hand only] 'Even though the animals', eh, hey 'animals' as well. Because they are talking about animal, right? [Line 7 continued, in phrase, left-hand only] 'animal is likely to' [P] [Line 7 again immediately, in part, left-hand only, with KWIC inserted] 'the animal is [E] likely to recover [E] recover the [P] [Option C] 'livestock''. I think it's [Option C] 'livestock', yeah. I think may sure from from this word? From the this word?

R: Uh-huh, [Line 4] 'slaughtering' and [P]

Dakota: 'Slaughtering' [from Line 4] and 'the animal' [from Line 7]. [P] And then 'illegal' [from Line 4] maybe. That's kind of thing I was thinking.

R: Uh, 'illegally' [from Line 4].

Dakota: And then 'stolen' [from Line 3], yeah. Something against [P] Yes, it's the weapons [from Line 2]. It's also similar to er 'slaughtering' [from Line 4] right? So, I think it's er

R: So, you can answer after after reading one to seven,

Dakota: Hmm,

R: Without having to go on to eight, nine and ten,

Dakota: Er, no, no,

R: Not at all?

Dakota: No.

R: Alright! Thank you so much.

Dakota: That's fine.

ConCloze 2: Esther on Item 5 (Target word: *livestock*)

Esther: [Option A] 'breeding', [Option B] 'creature creature', [Option C] 'livestock', and [Option D] 'stray'. Right. [P] the first one [Line 1] 'However, the droughts in nineteen sixty-nine to seventy-three and nineteen eighty four have considerably diminished dash [i.e., KWIC position] and pastures [P] thereby weakening some parental, elderly [E] and [E] chiefly leverage' [P] erm [P] OK, could be [Option C] 'livestock' also over here. Let see the next one. [P] [Line 2, in part] 'island laboratories, where they res where they search for new weapons against this [P] dash [i.e., KWIC position] disease [P] that threatens no consumer's life [in fact 'health']' [P] hmm, OK, then, could be the other one. [P] Erm [P]

R: Just say whatever you are thinking.

Esther: Again, I think [Option C] 'livestock' will be better over here.

R: For Line 3?

Esther: [Line 4, with KWIC inserted] 'available in the state food [E] markets [E] markets and anyone caught [E] slaughtering [E] [Option C] 'livestock' illegally [P] may be imprisoned for up to four years' [P] OK, [Line 5, with KWIC inserted] 'for the spread of sleeping sickness. There [P] [E] restocking [E] of [Option C] 'livestock' in Teso has been identified as a central activity' [P] OK, [Line 6, in part, with KWIC inserted] 'which can lead to [E] large increase [E] in [Option C] 'livestock' numbers? [P] [Line 7, in part, with KWIC inserted] 'Even though the animal is likely to recover [P] the [Option C] 'livestock' producer's worries [P] are hard are hardly over' Right. [Line 8, in part, left-hand only] 'the local lumber yard they could con contract to ship [P] [KWIC position here]' OK, [Line 8 again immediately, in part, with KWIC inserted] 'they could contract to ship their [Option C] 'livestock' to South St. Paul' Right? [P] [Line 9, in part, from right before the KWIC position, with KWIC inserted] 'while ranchers continue to [E] raise [E] [Option C] 'livestock'. [P] This approach has become increasingly popular' [P] er [P] [Line 10, in part, left-hand only, with KWIC inserted] 'escape with some of his [Option C] 'livestock'', yeah, it's [Option C] 'livestock', yeah.

R: Uh-huh. Are there any key words that you think that [E] help [E] you [E] specially [E] to get to the answer?

Esther: Because you see in this one, you're talking about [P] [Line 1, in word, left-hand only] 'droughts'

R: [Line 1, repeating] 'the droughts', uh-huh.

Esther: And you're talking about [Line 1, in word, right-hand only] 'pastures'

R: [Line 1, in word] 'pastures', uh

Esther: So, well, actually, if there is drought, well, well, there is a factor. Like if 'pastures' [from Line 1] are are diminishing.

R: Uh.

Esther: So, [P] and [Option C] 'livestock' is related [P] to it [i.e., Line 1]. [P] And then it's talking about [Line 1 again immediately, in part, right-hand only] 'weakening some parental, elderly and chiefly leverage' so like [P] kind of these things erm seem to be re related. And you see over here, [P] hmm 'slaughtering' [from Line 4]. So, when you slaughtering, what you have to slaughter is actually [IA] you can't slaughter a [Option A] 'breeding', you can't slaughter a [Option B] 'creature' [chuckle], so you have to slaughter a [Option C] 'livestock'

R: Hmm, interesting.

Esther: And then [Line 3, in phrase, from right before the KWIC position, with KWIC inserted] ‘stolen [Option C] ‘livestock’ from raids’ [P] ‘Stolen’, you can’t [E] steal [E] ‘breeding’ [Option A], or or you can’t steal [P] ‘stray’ [Option D] [P]

R: Uh, OK. That’s fine.

Esther: [IA] [chuckle]

R: Anything else that you would like to say about this question?

Esther: Hmm [P] because I understand what you’re saying, because, you know, sometimes more than two words can fit in the sentence.

R: Yeah.

Esther: But when you read more, I got, ev every sentence, there are a few cues. Again, because I think this sentence started, [P] I have the beginning of the sentence, so it [IA ‘could be’] easy for me to understand. And it’s

R: So so, Line 1 help you [IA ‘get rid’] yeah

Esther: Yeah, and I think it is the only complete sentence

R: Uh,

Esther: So, it became easier for me to [P] understand.

R: So, from from the first question [i.e., Line 1] to the fifth question [i.e., Line 5],

Esther: Uh-huh.

R: You seemed to er you seem to start from from the choices and then try to fit each into the sentences.

Esther: Yeah, trying to see what makes, which word will make sense over there.

R: Uh. What do you mean, ‘make sense’?

Esther: Because when you’re reading

R: Could you please explain more?

Esther: Because when you are reading a sentence [P], you are having a mental picture of [P] what, whosoever written in the sentence, whatsoever the author is trying to say,

R: Yeah.

Esther: So, it’s there the er what [IA ‘makes sense like’] in the blanks. [P] So out of these words [i.e., the options] you have to fit [IA], like you have to choose one of these words to fit in that space

R: Yeah.

Esther: So that the whole sentence becomes meaningful, it becomes like [P]. ‘Make sense’ means that when you [E] read it [E], you understand the meaning behind it.

R: Uh

Esther: You understand what what the author is trying to say.

R: Uh-huh.

Esther: And that’s why I [IA ‘was saying’] in the beginning if the beginning is not there, I don’t know what’s the subject, I don’t know how the sentence is constructed, and [P] the the end [IA ‘of those’] is not there, so it is somewhere [P] hanging in the middle

R: Yeah,

Esther: So, it's difficult to decide which word will fit there, [P] because er there could be more than one word. If if you know the beginning, if you know the end, [P] then you get the [E] complete [E] meaning. Like [IA], this is not the only erm, in this exercise, [E] this [E] one, the first one, [i.e., Line 1]

R: Uh-huh, Line 1.

Esther: The only complete sentence, so it is f it is much easier because it makes sense [P] what they are talking about. They are talking about the 'drought' [from Line 1] and this year, [P] and then er what 'diminished' [from Line 1] because [IA 'have considerably' [from Line 1]] 'the pastures' [from Line 1] and 'diminishing' [from Line 1] and then they had [E] an impact [P] on [E] the people [P] so it makes more sense.

R: Uh, OK. Lovely. Thank you.

ConCloze 3: Franz on Item 5 (Target word: *livestock*)

Franz: Haha. [laughing] [P] Easy guess? OK. erm [P] now, I'm translating [P] the answers first. [Option A] 'breeding', I know. [Option B] 'creature', [Option C] 'livestock', and [Option D] 'stray'. [P] [Option D] [E] 'stray' [E] [P] [Option D] 'stray'? [P] O, OK. I'm I'm confused erm between [Option D] 'stray' as an adjective and [Option D] 'stray' erm [P] as a noun.

R: Uh-huh.

Franz: So, let's see [P] what can be suitable for [P] erm the blanks provided here.

R: Yes.

Franz: Erm [P] [Line 1, in part] 'the droughts in tut tut tut [i.e., the year range in the original text] and nineteen [P] [E] nineteen eighty-four! [E] [P] [E] nineteen eighty-four! [E]' Come on, that sounds very [IA 'all real']! Haha [laughing] [P] It's so funny, 'nineteen eighty-four' [Line 1 continued] 'have considerably [E] diminished [E] [KWIC position here]' OK, I know, I know the word 'diminished' [from Line 1] And [P] 'decrease' [a paraphrasing going on?] I think, 'diminish', decrease [P] psst [P] [Line 1 continued] '[KWIC position here] and pastures thereby weakening some parental, elderly, and [E] chiefly [E] leverage' I don't know, I don't know the word, I don't know the meaning of 'leverage' [from Line 1]. [Line 2, in part] 'island laboratories where they [E] search [E] for new weapons against [E] this [E] [P] [KWIC position here] disease' hmm, psst, hmm, [P] s [P] OK, [P] OK, erm, the answer [P] erm, as I assume from two sentences I have read, I think the answer [E] should be [E] adjective or noun. I mean, the word itself could be both adjective and noun, because here, [P] I think here needs noun

R: Uh-huh

Franz: [Line 1, in phrase, from right before the KWIC position] 'diminished something [i.e., KWIC position] and something [the second constituent is 'pastures' in the real concordance line]' erm for second sentence it needs erm adjective,

R: Uh-huh.

Franz: So, let's take a look at the third sentence [i.e., Line 3].

R: Yeah.

Franz: [Line 3] 'trade allowed [P] Chiricahua [chuckle] Chiricahua, what! and the Apa and the other Apache' 'Apache''. That sounds very [P] familiar. It must be any group [P] or native Americans?

R: I don't know.

Franz: [chuckle] [Line 3 continued, from right before the KWIC position] 'to exchange their stolen [KWIC position here] from raids in Sonora and Chihuahua' Argh, [showing adoration] 'Chihuahua'! [from Line 3] 'Chihuahua'! 'Chihuahua'! [laughing]

R: OK.

Franz: Yes. [laughing] OK, [P] ‘Apache’, ‘Chi[ri]cahua’, ‘Chihuahua’ [P] So, ‘Chihuahua’ [all are the words from Line 3], oh, in this context, you have ‘guns’ [from Line 3] here, you have ‘food’ [from Line 3] here, [IA] er it must be related [E] to [E] ethnic group or native American.

R: Uh-huh.

Franz: [E] ‘slaughter’ [E] [from Line 4],

R: Uh-huh.

Franz: See?

R: Uh-huh.

Franz: Er, [Line 4, in part, right-hand only] ‘imprisoned [‘for up to’ not verbalized] four years’ [Line 5] ‘spread [‘of’ not verbalized?] sleeping sickness. The restocking [E] of [E] [KWIC position here] [P] Teso’ Er, it must be another ethnic group. [Line 5 continued] ‘has been identified as a central activity for tut tut tut’ [P] Hmm, [Line 6, in word] ‘localized’ [Line 7, in part, left-hand only] ‘walk. [P] Even though the animal is likely to recover, [E] the [E] [KWIC position here]’ [P] hmm, [P] hmm, psst, [P] I’m hesitating

R: Yeah.

Franz: Erm, [P] between [P] erm [Option A] ‘breeding’ and [Option C] ‘livestock’.

R: Yeah.

Franz: Erm, these two [Options A and C ‘breeding’ and ‘livestock’] [P] sound very relevant to each other, right? Erm but, [Option B] ‘creature’ doesn’t sound well. [Option B] ‘creature’ doesn’t sound fit to this context, and [Option D] [E] ‘stray’ [E] [P] no, no more [Option D] ‘stray’.

R: Uh-huh.

Franz: Erm so, [P] OK, [P] I I choose [Option C] ‘livestock’ [P]

R: Why?

Franz: [P] Honestly, [P] honestly, I I think the word ‘livestock’ [i.e., Option C] erm has a broadened meaning compared to [Option A] ‘breeding’. [Option A] ‘breeding’ can be only one angle of [Option C] ‘livestock’.

R: Uh-huh.

Franz: So, to be safe, [P] I choose [Option C] ‘livestock’

R: Uh-huh. [P] And, what about the contextual clues?

Franz: [P] Erm, [P] contextual

R: Yeah,

Franz: [P] Hmm, well, if you, if you choose the word [Option A] ‘breeding’ and place it [P] here [Line 8]

R: Sentence eight

Franz: Sentence number eight [Line 8]. It it doesn’t sound good. It sounds very weird, [Line 8, in phrase, right before the KWIC position, with KWIC inserted] ‘to ship their [Option A] ‘breeding’ What!

R: Uh-huh.

Franz: Right? [P] So it must be [Option C] ‘livestock’ apparently.

R: Alright.

Franz: I feel very [E] certain [E]

R: OK?

Franz: That my answer is correct.

R: Uh-huh.

Franz: OK?

R: OK. Anything else?

Franz: I [IA 'take a stand']. I hold a grip.

R: OK.

Franz: Haha [laughing].

R: Anything else that you would like to say about this item?

Franz: [P] Erm it's more difficult compared to the fourth one [i.e., Item 4].

R: Why?

Franz: Erm, because of [P] because of the strange [E] names [E]

R: Strange names?

Franz: 'Chihuahua', 'Sonora', 'Apache' [all from Line 3] [P]

R: Uh,

Franz: 'Chi Chiricahua' [all are the words from Line 3]. You you can be distracted by [P] by these strange names. And some erm other [P] vocabula vocabulary that you you do not comprehend its meaning. [P]

R: Uh-huh.

Franz: But however, [P] however, I [sigh] [P] psst, I can speculate, I think, from from, more or less from the con from the context to [P] to [E] cross [Option D] 'stray' and [Option B] 'creature' [E] out. So, I keep only [Option A] 'breeding' and [Option C] 'livestock', and I

R: Uh-huh.

Franz: Then I try to place erm these two words into [P] certain sentence, and I think [Option C] 'livestock', as I told you, [Option C] 'livestock' has broaden meaning, so it must be safer to choose this one.

R: So, again, why why did you first cross out [Option B] 'creature' and [Option D] 'stray'? Why?

Franz: Erm,

R: Why did you decide to cross out Choice B [Option B 'creature'] and D [Option D 'stray'] out?

Franz: Put it this way, put it this way, erm, er, from from the context, from the context of this group of sentences, I don't think [Option D] 'stray' [P] can be used, because if I if I'm correct, [Option D] 'stray' means jonjad [in his native language, which is the correct translation],

R: OK.

Franz: And [P] erm [P] this text is, this text talks about a [E] systematized erm farming [E] or something. Right? So, [P] er [probably from Line 6] the animal can't can't be [Option D] 'stray'.

R: OK.

Franz: If I if I'm right. If I'm correct. [IA] [Option D] 'stray'

R: Alright.

Franz: O, or it means fang [in his native language, which is the translation of 'straw'] hahaha [laughing]

R: I don't know. [IA] I cannot say.

Franz: O, OK. But but I, I cross this [Option D] 'stray', the word [Option D] 'stray' out.

R: Uh-huh, uh-huh. Because it [as the respondent has said all along] doesn't fit the line, the context at all.

Franz: Right, right. Erm, and for for [Option B] 'creature', fre [Option B] 'creature, creature' tends to mean [P] erm 'animal' and, in general, or 'animal' in [P] in broadest sense, I mean.

R: Why don't you choose it if it's, if its meaning is broad?

Franz: Because [P] put it this way, hey, come on, are you listening?

R: Yes, 'put it this way', yeah.

Franz: Because [Option B] 'creature' erm [P] [Option B] 'creature' means animal in general, erm [Option B] 'creature' [P] [Option B] 'creatures' erm can stand for wild animals?

R: For wild animals?

Franz: Wild animals. As as well as domesticating ones, right?

R: Yeah.

Franz: But, erm, the context erm doesn't convey erm the message about wild animals, more or less, or wild lives.

R: OK.

Franz: [E] at all [E]. So, I have more relevant [P] words to choose. Why do I [E] need [E] to keep [Option B] 'creature'? I cross it out. Does it make sense?

R: Right.

Franz: OK,

R: OK, thank you so much.

ConCloze 3: Gill on Item 5 (Target word: *livestock*)

Gill: The words are [Option A] 'breeding', [Option B] 'creature', [Option C] 'livestock', [Option D] 'stray'.
Er, [P] [Line 1, in part, with KWIC inserted] 'However, the droughts in 1969 to 73 and 1984 have considerably [E] diminished [E] [P] [Option A] 'breeding' and pastures', [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option C] 'livestock' and pastures' [P]

R: Uh-huh.

Gill: [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option D] 'stray' and pastures' [P]

R: Uh-huh.

Gill: [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[IA] [Option A] 'breeding' and pastures', [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option B] 'creature' and pastures', [Option B] 'creature creature' is out. [Option C] 'livestock', [P] [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option C] 'livestock' and pastures'. No, [Option C] 'livestock' is out. [Option D] 'stray', [Line 1 again immediately, in part, from right before the KWIC position, with KWIC inserted] '[IA 'diminished'] [Option D] 'stray' and pastures thereby weakening some parental, elderly, and chiefly leverage'. [P] Psst, OK, let me move to the second one [i.e., Line 2].

R: Alright.

Gill: [Line 2, in part, left-hand only, with KWIC inserted] 'island laboratories where they search for [P] new weapons against [E] this [E] s [KWIC position here]' [P] er [P] [Line 2 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'against thi this [Option C] 'livestock' [E] disease [E] that threatens' [P] [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'this [Option A] 'breeding' disease', no. [P] [Line 2 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'creature' disease', no. [Line 2 again immediately, in part, right-hand only, with KWIC inserted] '[Option D] 'stray' disease that threatens to con[su]mer's health' [P] threatens to no consumer's health' [P] [Line 2 again immediately, in part, right-hand only] 'that threatens no consumer's health, but every consumer's pocket[book].' [P]

[Line 2 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'stray, stray, stray, stray' disease' [P] [Line 2 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'stray' disease'. Psst, let me think more for it, and I'll [IA] here.

R: OK.

Gill: [IA] [Line 3, in part, left-hand only, with KWIC inserted] 'trade followed [in fact 'allowed'] Chiri Chiri Chiri Cari Cariki Charikihua [P] Charikihua [in fact 'Chiricahua'] [IA] and other Apache [P] to exchange their stolen [P] [Option C] 'livesto [livestock]'' [Line 3 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'stolen [Option C] 'livestock' from raids in' yeah [Line 3 again immediately, in word, right before the KWIC position, with KWIC inserted] 'stolen [P] [Option C] 'creature'' [P] [Line 3 again immediately, in word, right before the KWIC position, with KWIC inserted] 'stolen [Option D] 'stray'' [Line 3 again immediately, in word, right before the KWIC position, with KWIC inserted] 'stolen [Option C] 'livestock''. [P] OK, let me [IA] for it. [Line 4, in part, with KWIC inserted] 'available in the state food markets, and anyone caught [P] slaughtering [P] [Option C] 'livestock' illegally may be imprisoned' OK? OK, I will now I'm getting it clearer. So, [Line 4 again immediately, with KWIC inserted] 'for the spread of sleeping sickness. The restocking of [P] [E] [Option C] 'livestock' in Teso [E] has been identified as a central activity', OK, yeah. [Note that Line 5 appears not to be verbalized or skipped altogether here.] [Line 6, with KWIC inserted] 'Landau et al. which can lead to large increases in er [P] [Option C] 'livestock' numbers as a result to the occurrence of localized' [P] s [Line 7, in part, with KWIC inserted] 'Even, walk. Even though the animal is likely to recover [P] the [Option C] 'livestock' producer's [P] worries, producer's worries are hardly over' [P] [Line 7 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option A] 'breeding' producer's worries [IA]' uh-huh. Er er er [Line 8, in part, with KWIC inserted] 'the local lumber yard, OK, local lumber yard, they could contract to [E] ship [E] their [Option C] 'livestock' to [E] South [E]' [P] er [Line 9, in part, left-hand only] 'acquired through purchased conservation easements while', yeah, ' ranchers continue to raise [KWIC position here]' yeah, 'ranchers' [from Line 9] [IA 'has to do with'] [Option C] 'livestock'' [Line 9 again immediately, in phrase, left-hand only, with KWIC inserted] 'while ranchers continue to raise [Option C] 'livestock'', OK. [P] [Line 10, in part, with KWIC inserted] 'injure his family [P] and then escape with some of his [Option C] 'livestock' [IA 'When they return'] Farooq' yeah, [Option C] 'livestock'.

R: Uh-huh.

Gill: [IA 'It work like a'] 'ranchers' [from Line 9],

R: Uh-huh.

Gill: 'Ranchers' [from Line 9] will show, tell me like [Option C] 'livestock' the er [Line 8, in part, from right before the KWIC position] 'contract to ship, to ship' Yeah, here, 'to ship [KWIC position here] to St. Paul' The word 'ship' [IA] moving something. So, yeah, [Option C] 'livestock' is here. Then here too er, [Line 7, in part, left-hand only] 'even though the animal is likely to cov to recover, the [KWIC position here]'. You cannot [IA] [Line 7 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option A] 'breeding' producer's', cannot be [Line 7 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'creature' producer's', cannot er [Line 7 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'stray' producer's' OK. [The intention here seems to demonstrate, rather than to figure out the compatibility]

R: Uh-huh.

Gill: And then, here, [Line 6, in part, with KWIC inserted] 'which can lead to large increases in [P] [Option C] 'livestock' numbers', yeah. So, I'm going for [Option C] 'livestock'.

R: OK.

ConCloze 3: Halle on Item 5 (Target word: *livestock*)

Halle: [IA 'first one'] First line, er [P] [IA] [IA 'OK'] [Line 1] 'However the droughts in nineteen sixty-nine seventy-three and nineteen ninety-four [in fact '1984'] have considerably diminished [E] blank [E]

[i.e., KWIC position] and pastures, thereby weakening some parental, elderly, and chiefly lever leverage' And the words are [P] [E] [Option A] 'breeding', [Option B] 'creature', [Option C] 'livestock', and [Option D] 'stray' [E]. [P] So, [Line 1 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'have considerably diminished [Option A] 'breeding' and [E] pastures [E]' [P] Argh, I don't think it's that one. [Line 1 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'have considerably diminished [E] [Option B] 'creature' [E] and pastures' Again, hmm, doesn't really work. [Option C] 'livestock' though, makes sense, [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option C] 'livestock' and pastures'? Or [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option D] 'stray' and pastures'? No, not really. I'm gonna go with [Option C] 'livestock' for now. I'll see next one. So, second one, [Line 2] 'island laboratories where they search for new weapons against this er [P] blank [i.e., KWIC position] disease that threatens no consumer's health, but every consumer's pocketbook' [P] Well, I think [P] [Line 2 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option A] 'breeding' disease'? [P] Na, uh-er, [P] [Line 2 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'creature' disease', no. Er, I think [Option C] 'livestock' would work here. [Line 2 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'livestock' disease', would work, I think it would work again. [P] So, let's see the third one, [Line 3, with KWIC inserted] 'trade allowed Chiricahua and other Apa Apache to exchange their [E] stolen [E] [P] [Option C] 'livestock' from raids in Sonora and Chihuahua for the food, guns' [P] Definitely, '[Option C] 'livestock', [Line 3 again immediately, in word, right before the KWIC position, with KWIC inserted] 'stolen [Option C] 'livestock''. [P] Does it make sense? Or the other ones? That's good coz it'll give me a clue, and then, fourth line, [Line 4] 'available in the state food markets, and anyone caught [E] slaughtering [E] [P] blank [i.e., KWIC position] illegally may be imprisoned for up to four years, but' Again, [Line 4 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'slaughtering the [Option C] 'livestock'' makes sense, works. So, the fifth one, [Line 5] 'for the spread of sleeping sickness. Four, the restocking of [P] blank [i.e., KWIC position] in Teso has been identified as a trenal central activity for' So, [Line 5 again immediately, in phrase, right before the KWIC position, with KWIC inserted] 'restocking [E] of [E] [Option C] 'livestock'' makes sense. [P] again [chuckle] [IA 'have'] 'restocking' [from Line 5] [P] So,

R: Uh-huh.

Halle: It has the obvious link there, [P] then the sixth line [Line 6] 'Landau et al. 1995 which can lead to large increases in blank [i.e., KWIC position] numbers and, as a result, to the occurrence of localized' again [Line 6 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option C] 'livestock' numbers' works. Then seventh line, oh, I see there are ten here [i.e., ten concordance lines, rather than seven], [Line 7] 'walk. Even though the animal is likely to [E] recover [E], the blank [i.e., KWIC position] producer's worries are hardly over, notes microbiologist Cecelia A. Whetstone, head' [P] [Line 7 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] '[E] the [E] [sigh] [P] [Option C] 'livestock' producer's'. Again, just to make sure, [Line 7 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option A] 'breeding' producer's' [E] might [E] work, but it's a bit strange. [Line 7 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option B] 'creature' producer's' just weird. And [Line 7 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'stray' producer's'? [IA] Well, so, I'm gonna choose [Option C] 'livestock'. Then eighth, [Line 8] 'the local lumber [E] yard, [E] they could contract to ship their [P] blank [i.e., KWIC position] to South St. Paul, and they could [E] deal [E] with local grain' Again, [Option C] 'livestock' works really well. Should be [Option C] 'livestock'.

R: Uh-huh.

Halle: Nine, [Line 9] 'acquired through purchased con conservation easements while ranchers continue to raise blank [i.e., KWIC position]. This approach has become increasingly popular, especially among NGOs such' [P] [Line 9 again immediately, in word, right before the KWIC position] [E] 'raise' [this word is right before the KWIC position] [E], 'ranchers, ranchers' [this word is three words before the KWIC position; the first noun before the KWIC position], [Line 9 again immediately, in phrase, right

before the KWIC position, with KWIC inserted] ‘continue to raise [P] [Option C] ‘livestock’’ That makes sense in the context. And finally, [Line 10] ‘injure his family, and then escape with some of his [P] blank [i.e., KWIC position]. When they return, Farooq has gathered friends to help him’ [Line 10 again immediately, in part, left-hand only, with KWIC inserted] ‘injure his family, and then escape with some of his [Option C] ‘livestock’’ again, work best. [P] So, I’m gonna say [Option C] ‘livestock’.

R: OK.

Halle: Should I click?

R: Yeah, you can continue.

ConCloze 3: Igor on Item 5 (Target word: *livestock*)

Igor: Psst, OK.

R: Keep saying.

Igor: [IA]

R: Keep saying whatever comes to your mind.

Igor: [IA] [Line 1, in phrase, right before the KWIC position] ‘have considerably diminished [KWIC position here]’

R: Keep saying whatever you are thinking.

Igor: [Line 1 again immediately, in phrase, right before the KWIC position, with KWIC inserted] ‘diminished and [P] [IA] [Option A] ‘breeding’’? It’s [IA ‘meaning’] [Line 1 again immediately, in phrase, right after the KWIC position, with KWIC inserted] ‘[Option A] ‘breeding’ and pastures’. [Line 1 again immediately, in part, from right before the KWIC position, with KWIC inserted] ‘diminished [Option A] ‘breeding’? [IA] parental, elderly and chief[ly] leverage’ [P]

R: Keep saying.

Igor: [Chuckle]

R: Please.

Igor: [Line 2, in part, left-hand only] ‘island laboratories where they search for the new weapons [in fact ‘for new weapons’] against this [KWIC position here]’ [P] [Line 2 again immediately, in part, from right before the KWIC position] ‘against this [KWIC position here] [E] disease [E]’ [IA ‘‘new weapons’ and ‘disease’, I don’t know’] [P] ‘[KWIC position here] disease [P] [IA ‘that threatens no consumer’s health’] consumer’s health [IA]’

R: Keep saying.

Igor: Psst, hmm, [P] [Option D] ‘stray’, [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] ‘this [Option D] ‘stray’ disease’ [P] [Line 2 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] ‘this [IA] [Option C] ‘livestock’ disease’ [P]

R: Keep saying.

Igor: [P]

R: Please.

Igor: Haha [laughing]

R: I would like to know what what you are thinking.

Igor: Yeah, [chuckle] [IA] Give time for me, just

R: Yeah, just say it what [IA]

Igor: When I, when I got conclusion, I will explain to you.

R: It's OK. Just just say what whatever words that is, you know,

Igor: No, we cannot say before we think. Yeah, we have to think first.

R: You can do that the same time, of course.

Igor: When I say, I I cannot [P] thinking [IA].

R: Please. Say and think at the same time. No worries about your grammar or

Igor: We need to concentrate. We

R: The structure, or words.

Igor: [Line 3, in phrase, right before the KWIC position, with KWIC inserted] 'exchange their stolen [KWIC position here]'

R: No worries whether your speech is broken or not. Just say it.

Igor: [Line 3 again immediately] 'trade allowed ['Chiricahua' not verbalized] and other ['Apache' not verbalized] to exchange their stolen [KWIC position here]' [P]

R: Keep saying.

Igor: [Line 3 continued] '['from raids in Sonora and Chihuahua' not verbalized] for the food [IA 'guns'] OK, now, [Option A] 'breeding' is not correct, I think, because, when I apply it, [Line 3 again immediately, in phrase, from right before the KWIC position, with KWIC inserted] 'their stolen [Option A] 'breeding' from' No. It's not correct. [P] Hmm, [Line 3 again immediately, in word, right before the KWIC position, with KWIC inserted] 'stolen [P] [Option C] 'livestock'' [P] [Accidental snap] Sorry.

R: It's alright.

Igor: [Line 3 continued] '[KWIC position here] from raids' [P] [Line 4, in part] 'available in the state food markets, [DISTRACTION] [P] and anyone caught [P] slaughtering [P] [KWIC position here] [ille]gally' [P] Hmm, [P] [Option C] 'creat[ure]' or [Option D] 'livestock'? [P] [Line 5, with KWIC inserted] 'for the spend [in fact 'spread'] of sleeping sickness. The [P] restock[ing] of [P] restock[ing] of [Option C] 'livestock' in Tesco [in fact 'Teso'] has been identified as central activity' [P] Psst, [IA 'What's'] [Option D] 'stray stray', I don't know the meaning. [chuckle] [P]

R: It's OK.

Igor: [Line 6, in part, with KWIC inserted] 'which can lead to a large increase [in fact 'large increases'] in [P] [Option D] 'stray' numbers' [P] [Line 7, in part, with KWIC inserted] 'walk. Even though ['the' not verbalized] animal is likely to recover, [E] the [E] [P] [Option B] 'cr creature' produce[r]'s' worries' [P] [Line 7 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option C] 'livestock' produce[r]'s' worries', no, psst. [P] [Line 8, in part, left-hand only, with KWIC inserted] 'the local lumber yard, they could contract to ship [E] their [E] [P] their [Option D] 'stray' [P] s' [P] Now I like er [P] [Option C] 'livestock' and [Option D] 'stray'. But I don't know which word should [P] appropriate on here.

R: Uh-huh.

Igor: Hmm [P] [Line 1, in phrase, from right before the KWIC position] 'diminished [KWIC position here]' [P] [Line 1 again immediately, in phrase, from right before the KWIC position] 'considerably diminished [KWIC position here]' [P]

R: Keep saying.

Igor: [Line 1 continued] '[KWIC position here] and pastures' [P] [Line 10, in part, left-hand only, with KWIC inserted] 'injure his family? And then escape with some of [E] his [E] [P] his [P] [Option D] 'stray'' [Line 10 again immediately, in part, from right before the KWIC position, with KWIC inserted] 'h[is] [Option C] 'livestock'? [P] When they return, Farooq has gathered friends to help him' and [P]

[Line 10 again immediately, in part, left-hand only] 'injure his family, and then escape with some of his' [P]

R: Keep saying.

Igor: [Line 10 continued] 'his er [KWIC position here]' [P] OK, I I choose er [Option C] 'livestock', yeah.

R: Why?

Igor: Psst, [P] because er when I apply it with er [Line 10 again immediately, in phrase, right before the KWIC position] 'some of his [KWIC position here] [IA 'phrase']', should be 'his friend[s]' or something like this [referring to a position of a noun phrase? OR a concrete animate noun?], I think. So the the meaning should match with

R: What do you mean, the the meaning, the meaning [E] of Choice C [Option C 'livestock'] [E] matches

Igor: Yeah, match.

R: Line 10?

Igor: Yeah.

R: In what way?

Igor: Yeah, yeah, yeah.

R: Could you please explain?

Igor: Hmm, [P] [Line 8, in phrase, left-hand only] 'a [in fact 'the'] local lumber yard'. I think er the Choice C [Option C 'livestock'] can can use with er line number 10? [i.e., Line 10] [P]

R: Uh-huh.

Igor: And, [P] the line number [P] er, let me see [P] I line number 5? [i.e., Line 5], yeah, [Line 5, in phrase, from right before the KWIC position, with KWIC inserted] 'restock[ing] of [Option C] 'livestock' in Tesco [in fact 'Teso'] has been identified'. But I think the [Option A] 'breeding' and [Option B] 'creature' is not [P] is not appropriate. But I I don't know the

R: To which

Igor: Meaning of the

R: Line?

Igor: Hmm, [P] to the line number 10, [P] [Line 10, in phrase, right before the KWIC position, with KWIC inserted] 'of of his [Option A] 'breeding'', I don't think. [IA] It it [Line 10 again immediately] 'injure his family [P] and then escape with some of his [KWIC position here]' [P] psst, [Line 10 continued] 'When they return, Farooq [chuckle] has gathered friends to help him' [P] Psst, well [Option A] 'breeding' [chuckle]. [Line 10 again immediately, in part, left-hand only, with KWIC inserted] 'injure [several words not verbalized here] and [Option A] 'breeding'' and this one about [P] health [potentially Line 2] [P] [Line 1, in part, right-hand only] '[KWIC position here] weakness [in fact 'weakening'] some parent[al], elderly' [P] Hmm, [Line 2, in part, with KWIC inserted] 'they search for new weapons against this [Option A] breeding disease' haha [laughing] [P] It's difficult for me, this one.

R: No worries.

Igor: OK, yeah, just [P] guess. I I guess [IA] [Option C] 'livestock'

R: Uh-huh. [P] So, which line which line helps you most to [P] to come to the conclusion that it should be Choice C that is the most appropriate one?

Igor: Er I

R: Which line?

Igor: I think [P] number four [i.e., Line 4]

R: Four?

Igor: Yeah.

R: What are [P] its key words that help you most?

Igor: Er, it's about, talking about food and something about, yeah, should talking about not not [Option A] 'breeding', this one [i.e., according to his interpretation, Line 4 does not have anything to do with Option A 'breeding']. [NB: A part from Line 4: 'anyone caught slaughtering [Option C 'livestock'] illegally'] And not [Option B] 'creature'.

R: Uh-huh.

Igor: Yeah, I decide to choose erm [Option C] 'livestock' from [E] here [E]

R: Alright.

Igor: [Line 5, in part, with KWIC inserted] 'for the spree [in fact 'spread'] of [E] sleeping sickness [E]. The restocking of [Option C] 'livestock' in Tesco in Teso has been identified as a central activity' But all of them can apply, yeah.

R: Uh-huh.

Igor: This one I choose C [i.e., Option C 'livestock'] [P]

R: OK.

Igor: Guess. I did not know much [P] vo vocab.

ConCloze 4: James on Item 5 (Target word: *livestock*)

James: [P] Yeah, [Option A] 'breeding', [Option B] 'creature', [Option C] 'livestock', OK. [P] [Line 1] 'However, the [P] droughts in nineteen sixty-nine and eighty-four [in fact '1984'] have considerably dimi[P]nished and blank [i.e., KWIC position] [in fact the KWIC blank and then 'and'] pastures thereby weakening some [P] parental, elderly, and chiefly leverage.' [Line 2] 'island laboratories where they search for new weapons against this [KWIC position here] [P] disease [P] er ['that' not verbalized] threatens no consumer's health, but every consumer [P] pocketbook' [Line 3] 'travel [in fact 'trade'] allowed [P] [IA 'what is this?'] Chikwi chihua [i.e., Chiricahua] and other Apache to exchange their stolen [P] fr blank [i.e., KWIC position] from raids in [P] Sono[ra] and Chihuahua for food, [P] guns' [Line 4] 'available in the state food markets, and anyone caught [P] [IA 'what is this?'] slaughtering [P] blank [i.e., KWIC position] illegally may be imprisoned for up to four years, but' [Line 5] 'for the spread of sleep[ing] sickness. [P] four, ['the' not verbalized] restocking of [P] blank [i.e., KWIC position] in [P] Teso has been identified as a central activity for' [P] [Line 6] It's a reference [in fact '(Landau et al. 1995)'] 'which can lead to large increase in blank [i.e., KWIC position] numbers and, as a result, to the occurrence of localized' [Line 7] 'walk. [P] Even though the animal is likely to recover blank [i.e., KWIC position] producer's worries are hardly over, notes microbiologist [E] Cecelia [E] [P] A. Whetstone, head' [Line 8] 'the local [P] lumber yard, they could contract to ship [E] their [E] blank [i.e., KWIC position] to St. to South St. Paul, and they could deal with local grain' [P] OK, er [Line 9] 'acquired through purchased conservation es easements while ranchers continue to rise [in fact 'raise'] [KWIC position here]. [P] This approach has been[?] [in fact 'become'] increasingly popular, especially among NGOs such' er [Line 10] 'injury [in fact 'injure'] his family, [P] and then escape with some of this blank [i.e., KWIC position] When they return, Farooq has gathered friends to help him' So, [Item instructions] 'All [E] the [E] lines above miss the same word. Which of the following should be that word?' OK? And words, [Option A] 'breeding', [Option B] 'crea [P] creature', [Option C] 'livestock', and [Option D] 'st stray'. [P] [Line 1, in part, with KWIC inserted] 'However, [IA] and nineteen [P] eighty-four have considerably diminished er [Option B] 'creature' and pastures thereby weakening [IA]' [P] [Line 2, in part, with KWIC inserted] 'island laboratories where they search for new weapons against this [P] [Option B] 'creature' disease' [P] [Line 2 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option B] 'creature' disease that threats [in fact 'threatens']' OK. [P] [Line 3, in part, left-hand only] 'trade ['allowed Chiricahua' not verbalized] and other Apache to exchange their stolen [KWIC position here]' yeah [P]

R: Keep saying.

James: [Line 8, in part, with KWIC inserted] ‘the local lumber yard, they could contract to ship their [P] er [P] [Option A] [E] ‘breeding’ [E] to South Paul, and’ [P] [Option B] ‘creature’ [Option A] ‘breeding’ [P] [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] ‘diminished [Option A] ‘breeding’ and pastures’ er, [Line 3, in word, left-hand only] ‘Apache’ and s [Line 5, in part, left-hand only] ‘spread sleeping sickness. Four, livestocking [should be ‘restocking’] [KWIC position here]’ [P] and [Line 10, in part, with KWIC inserted] ‘injure his family, and escape with some of his [P] [Option A] ‘breeding’. While [in fact ‘When’] they return’ [P] I think er [P] I’m not sure this one, how, I can [P] apply the same word for all of this?

R: Yes, sure!

James: [chuckle]

R: Sure.

James: [IA]

R: I created the item, I am sure. [chuckle]

James: [Line 1, in part, from right before the KWIC position, with KWIC inserted] ‘considerably diminished [Option A] ‘breeding’ and pastures, thereby weakening some parental’ [P] I think er the most suitable one is er [Option A] ‘breeding’. This, [Option A] ‘breeding’, Answer A.

R: Uh-huh.

James: Because er er same as the previous recently, most of the, most of these blanks er, I think, er, [P] good with [Option A] ‘breeding’ [i.e., the selected KWIC fits most of the concordance lines given]

R: Uh-huh.

James: The word [Option A] ‘breeding’, the key word [Option A] ‘breeding’ [i.e., the best option to James is [Option A] ‘breeding’]

R: Uh-huh.

James: So, can I choose?

R: And what, what are the key words in these ten lines, that you think, maybe matches

James: Er

R: Or, yeah, compatible with the the br the Choice [Option A] ‘breeding’?

James: Like er, this one, [Line 8, in part, left-hand only] ‘the local lumber yard, they could contract to ship their [KWIC position here]’ you know, this one

R: Which one?

James: To [Option A] ‘breeding’.

R: Uh-huh.

James: [Line 8 again immediately, in phrase, from right before the KWIC position] ‘To [‘ship’ not verbalized] their [KWIC position here] to South Paul’

R: I mean, which words in that line?

James: Er, [Line 8 again immediately, in phrase, right before the KWIC position] ‘[E] ship [E] [P] their [KWIC position here]’, the [Line 8 again immediately, in phrase, right before the KWIC position] ‘ship their [KWIC position here]’ that one, that two words, the

R: Uh-huh.

James: And [Line 9, in word, right before the KWIC position] ‘raise [KWIC position here]’ the, [P] the, very er words, very close to this blank, it’s making good sense to this one, [Option A] ‘breeding’

R: Uh-huh.

James: And the, so [Option C] ‘livestock’, it’s kind of er [IA ‘meat’] or something. I think er it’s sometimes er

R: OK,

James: So, er I think the first one also talking about some meaning not the [Option C] ‘livestock’

R: Uh-huh.

James: So, I think this one.

R: OK, yeah.

ConCloze 4: Klavier on Item 5 (Target word: *livestock*)

R: [IA] you’re thinking.

Klavier: Well, er, [Line 1] ‘However, the droughts in nineteen sixty sixty-six and seventy-three and nineteen sixty-four [‘have’ not verbalized] considerably diminished [P] er [KWIC position here] and pastures, thereby weakening some parental, elderly, and chiefly leverage’ [P] [Line 2] ‘island laboratories where they search for new weapons against this [P] [KWIC position here] disease, that threatens no consumer’s health, but every consumer’s pocketbook’ [P] [Line 3, in part] ‘trade allowed Chi Chirica’ haha [laughing] can’t pronounce

R: Yeah, no worries.

Klavier: Yeah, [Line 3 continued] ‘and other Apache to exchange their stolen [P] [KWIC position here] from raids in Sonora’ [P] erm [P] I think it’s [Option C] [E] ‘livestock’ [E] erm because it’s mainly about animals [content from Line 1], and foods [word from Line 3], and, as I’ve just quickly read through the rest of the [P] sentences, erm [P] and [Option C] ‘livestock’ is [P] animals, erm like as animals as a product

R: Which line is about, could you said, [IA ‘fourth’]? Which line?

Klavier: Erm [P] Well, there was erm, er ‘food markets’ [from Line 4], ‘slaughtering’ [from Line 4], ‘their stolen’ [from Line 3] [P] blah blah blah, so just like ‘food’ and ‘guns’ [from Line 3], so that’s [P] so, it’s [Option C] ‘livestock’ as a products erm [P] [Line 2, in word, right after the KWIC position, with KWIC inserted] ‘[Option C] ‘livestock’ disease’ erm [P] [Line 5, in word, right before the KWIC position, with KWIC inserted] ‘restocking [‘of’ not verbalized] [Option C] ‘livestock’

R: Uh-huh.

Klavier: Erm [P] [Line 7, in part, with KWIC inserted] ‘Even though the animal is likely to recover, the [Option C] ‘livestock’ producer’s worries are [P] hardly over’ erm, yeah, [IA ‘that’s all’].

R: Alright, OK.

ConCloze 4: Lulu on Item 5 (Target word: *livestock*)

Lulu: The last one, great! [Option A] ‘breeding’, [Option B] ‘creature’, [Option C] ‘livestock’, [Option D] ‘stray’. Hmm, [P] hmm, OK, [Line 1, in part, with KWIC inserted] ‘However, the droughts in nineteen sixty-nine to seventy-three and have considerably [E] diminished [E] [P] [Option C] ‘livestock’ and pastures, thereby weakening some [P] parental, elderly, chief [IA]’ [Line 1 again immediately, in word, right before the KWIC position] ‘diminished er [KWIC position here]’ [P] hmm, psst, I’m I’m not sure, [P] something. My first instinct, it was [Option C] ‘livestock’, but but, I don’t know. So, the second one, [Line 2, in part, left-hand only] ‘island laboratories, where they search for new weapons [P] against [E] this [E] [KWIC position here]’ [P] [Line 2 again immediately, in part, from right before the KWIC position] ‘against this [KWIC position here] disease that threatens no consumer’s [P] health, but every consumer’s [P] pocket [in fact ‘pocketbook’]’ [P] [Line 2 again immediately, in part, left-hand only, with KWIC inserted] ‘they search for new weapons against this [P] hmm, [Option C] ‘livestock’

[Option B] 'creature', [Option A] 'breeding', [Option C] 'livestock' [P] Psst, still feel like it's [Option C] 'livestock' [P] like it's er [P] it's like it's erm [P] psst, contrasted to [P] [E] to [E] [P] it's something that's [E] consumed [E] [P] something that's consumed [P] psst, OK, consumed [Option C] 'livestock', can you? You eat beef, lamb, [P] but you don't eat [Option A] 'breeding'. Or you don't, [Option B] 'creature' is too [E] broad [E]. [Option D] 'stray'? [P] I I don't even think that [Option D] 'stray' fits in this context. [P] Hmm, [Line 3, in part] 'trade allowed er [P] [E] Chiricahua [E] and [E] other [E] Apache to exchange their stolen [P] [KWIC position here] from raids' Psst, [P] er, I think in this context, normally people say they steal [Option C] 'livestock'. It's er domestic animal [IA] to be raised by family, perhaps. Erm, [P] yeah, I think [Option C] [E] 'livestock' [E] [P] I'll go for [Option C] 'livestock'. Psst, [Line 10, in part, left-hand only, with KWIC inserted] 'injure his family, and then escape with some of his [Option C] 'livestock''. Yeah. [Option C] 'livestock', they [IA 'in live with'] his family and they raise family. So, it makes sense that if they escape, and then [P] er they take the [Option C] 'livestock' with them. Like in a movie, erm, psst, [IA 'life as a pie, yeah'] When they left India, they took er their animals into the zoo with them. Haha [laughing]

R: [Chuckle]

Lulu: So, yeah, I'll go for this one.

R: [Chuckle] I've never seen that film. OK, perfect. Thank you.

Lulu: [Chuckle]

Appendix 8. Verbal Reports for Modified Constructed-response Processing Analyses in ConCloze 7 (Section 4.3.4)

ConCloze 7: Maya on Item 1 (Target word: *social*)

R: Keep saying.

Maya: O, OK. [P] One, [Line 1] ‘which the categories of Islamic anthropology and something [i.e., KWIC position] and analyt analysis are to be operated and investigated’ [P] Two, [Line 2] ‘other sexual and drug use behaviors and [P] something [i.e., KWIC position] and demographic characteristics, we aimed to further’ [Line 3] ‘brightest and best students [P] a sense of something [i.e., KWIC position] and intellectual responsibility. [P] It’s [P] able to’ Four [Line 4] ‘of the Christian community and the [E] subsequent [E] something [i.e., KWIC position] and political expectations which were raised. [P] His’ Five [Line 5] ‘of the three subscales, Personal Behavior Difficulty [P] something [i.e., KWIC position here] Behavior Difficulty and Emotional Difficulty indicated’ [P] Six, [Line 6] ‘context of im of male family members’ encouragement, women’s something [i.e., KWIC position] pressure [P] and [‘positive’ not verbalized] [IA ‘reinforcement’] of praise’ [P] [Line 7] Seven ‘human society shaped by culture. In fact, something [i.e., KWIC position] scientists and historians are much more prone’ [P] [Question stem] ‘All the lines above miss the same word. [E] What should be that word? [E]’ [P] Hmm

R: Keep saying whatever you are thinking.

Maya: Erm, I don’t know actually. I dunno what this word should be. [P] Hmm, [IA] and I’ll have to read it again. [P] Hmm [Line 1, in part, left-hand only, with KWIC inserted] ‘which the categories of Islamic anthropology [E] and [E] [P] ‘the’ [i.e., KWIC position]? [P] ‘the’ is the word. [Maya wrote the word ‘the’ in the KWIC blank of Line 1, i.e., her response constructed ‘the’ filled on the paper.] Something so simple. [Line 1 again immediately, in part, right-hand only] ‘[KWIC position here] [IA ‘analysis’] are to be sort of operated and investigated’ [Line 2, in part, left-hand only] ‘other sexual and drug use behaviors, and [KWIC position here]’ [Maya was going to start writing the word ‘the’ (constructed) in Line 2] Ah, no. Wouldn’t go in, that word. [chuckling] [P] [Line 2 again immediately, in phrase, from right before the KWIC position] ‘and something [i.e., KWIC position] and dem’ [Line 2 again immediately, in word, right after the KWIC position] ‘something [i.e., KWIC position] [‘and’ not verbalized here] demographic’ [Line 2 again immediately, in word, right before the KWIC position] ‘behaviors [‘and’ not verbalized] [KWIC position here]’ [P]

R: Keep saying. Whatever you are thinking.

Maya: Erm, I don’t know that question [i.e., the concordance line]. [P] So, it can’t it actually can’t be the the word the same words. [P] Erm. [P] Actually dunno. [P]

R: OK. Now, if you have four choices here [the answer options now placed before Maya],

Maya: Erm. [P] [Looking at the answer sheet.]

R: Keep saying. Whatever you are thinking or reading.

Maya: Erm. [P] [Option A] ‘A common’, [Option B] ‘economic’, [Option C] ‘important’, [Option D] ‘social’. [P]

R: Keep saying.

Maya: [IA] [P] Ow. [Option D] ‘S social’. It’s [Option D] ‘social’. [Maya was writing the word in the first KWIC blank.] Because [P] of erm [the phrase ‘Islamic anthropology’ in Line 1 being underlined] [Line 1, in word, left-hand only] ‘Islamic’ [from Line 1] it’s like [Option D] ‘s social’ [P] erm [P] in this [Option D] ‘social’ context [P] [Line 2, in part] ‘drug and use [P] [NB: Words in Line 2 but rearranged partly] sexual and drug use behaviors and [IA ‘demogr charac characteristic’]’ That’s something to do with [Option D] ‘social’ behavior [NB: Not the exact word sequence presented in text] Erm [Line 3, in part, with KWIC inserted] ‘brightest and best students a sense of [P] [Option D] ‘social’ and intellectual responsibility.’ Make sense. Erm. [P] [Line 4, in phrase, right after the KWIC position] ‘political

expectations' [being underlined] again. And [Line 4 again immediately, in phrase, left-hand only] 'the Christian community' [being underlined]. Again, has something in relation to [Option D] 'social'. [P] erm [P] [Line 5, in phrase, left-hand only] 'of the three subscales,' Again, [Line 5 again immediately, in phrase, right before the KWIC position] 'Personal Behavior' ['Personal Behavior Difficulty' being underlined] [P] and er [Line 5 again immediately, in phrase, right before the KWIC position] '[IA 'the behavior'] 'Difficulty and Emotional' ['Behavior Difficulty' and 'Emotional Behavior Difficulty' being underlined]' Again, related to the word [Option D] 'social' [P] [Line 6, in part, left-hand only] 'context of male family members'. Again, [Line 6 again immediately, in phrase, left-hand only] 'family [P] members' ['family members' being underlined]' [Line 6 again immediately, in word, left-hand only] 'women's' [being underlined] [P] [Line 6 again immediately, in word, right after the KWIC position, with KWIC inserted] '[Option D] 'social' pressure' [blank KWIC filled in with the chosen word] er, that makes sense. Erm [P] [Line 7, in phrase, left-hand only] 'human society' [being underlined] [P] er [Line 7 again immediately, in phrase, right after the KWIC position, with KWIC inserted] '[Option D] 'social' scientists ['and' not verbalized here] historians' [KWIC blank filled out] [the words 'scientists' and 'historians' being underlined] [P] [the pen is put down]

R: OK. So, at first you did not have the choices, so what what happened?

Maya: Erm, it was really hard to pick a word that would go with all of them. Erm so initially, my first thought was that er it was connecting word [probably referring to a function word such as a coordinating conjunction] rather than a particular this verb [in fact 'social' is an adjective] to describe the words [P]

R: Uh-huh. But when you were given four choices, what happened?

Maya: Erm, I scanned through all the words and thought which one would seem appropriate erm and then I looked at all the sentences and picked out the. For example, I highlighted the 'Islamic anthropology' [from Line 1], 'Christian community' [from Line 4] [the phrases were pointed at] so they're er associated with people, and the word [Option D] 'social' means to have demographic backgrounds er people erm and in terms of er culture as well. So, this the word has more than one association.

R: OK, anything else you would like to add.

Maya: Nope.

ConCloze 7: Nina on Item 1 (Target word: *social*)

R: Yeah, sure. Keep saying.

Nina: [Line 1] 'which the categories of Islamic anthropology and [KWIC position here] [P] something analysis [IA 'are to be operated and'] investigated' [Line 2] 'other sexual and drug use behaviors, and [KWIC position here] [IA] demographic characteristics we aimed to further' [P] [Line 3] 'brightest [P] and best students a sense of [KWIC position here] [P] intellectual responsibility. It is able to' [P] [Line 4] 'of Christian community and subsequent [KWIC position here] and political expectations [P] which are raised. His' something [P] [Line 5] 'of the three subscales, Personal Behavior Difficulty [KWIC position here] [P] Behavior Difficulty and [IA 'Emotional Behavior Difficulty, indicated']' [P] [Line 6] 'context of male [P] male family members' encouragement woman's [P] something [i.e., KWIC position] pressure [P] [Nina's pen was still moved along the concordance line]' [P] Hmm.

R: Keep saying.

Nina: [Chuckling] [Line 7] 'Human society shaped by culture. In fact, [KWIC position here] [IA 'scientists'] [Nina's pen was still moved along the line to the end]' [P] Hmm [Line 7 again immediately] '[KWIC position here] scientists and historians are much more prone' [P] [Item instructions] 'All the lines above miss the same word. What sound [P] what should be that word?' [P] OK. [P] So, we have [Line 1, in part] 'categories [the word being underlined] and [KWIC position here] [P] and [IA] [a large part of the line was not said out loud] to be operated and investigated' [P] Hmm. [Line 5 pointed at] And this is something. Personal is not. Subscales? [P] Personal, [P] Emotional [NB: there was a parallel structure after 'the three subscales' in Line 5] something with [P] the mind? Maybe? [P] [Line 2, in part, left-hand only, with KWIC inserted] 'other sexual and drug ['use' not verbalized] behavior and [KWIC

position here] [P] ‘ethnic’ maybe. [P] No, doesn’t make any sense. [chuckling] [Line 3, in part, left-hand only] ‘brightest and best students a sense of [KWIC position here] [IA]’ [P]

R: Keep saying.

Nina: [Chuckling] [Line 4, in part] ‘the Christian community [IA] subsequent [KWIC position here]’ [P] Well, I can’t say it, the same word. Doesn’t make any sense. [P] [Line 6] context of male family members’ encouragement, women’s [KWIC position here] [P] pressure [IA] [Nina’s pen was moved along the rest of the line] of praise’ [Line 7] ‘human society shaped by culture. In fact, [KWIC position here] [P] scientists and [IA]’ [P] Hmm, [P] I I have no idea. Haha [laughing] I have no idea. [P]

R: Alright. So,

Nina: So difficult. [chuckling]

R: What if I give you four choices? [The option sheet placed before Nina, under the question sheet]

Nina: That that’s better.

R: [Chuckling] OK, think aloud, think aloud.

Nina: OK, the words are [Option A] ‘common’, [Option B] ‘economic’, [Option C] ‘important’, and [Option D] ‘social’. [P] [Option D] ‘social’ may fit. [P] It fits well here [the word ‘Behavior’ in Line 5 being underlined] because of the ‘behavior’ [from Line 5]. It has ‘Personal’ [from Line 5] and ‘Emotional’ [from Line 5], so. [P] Yeah, [Option B] ‘economic’ doesn’t fit well really well. [P] [Line 6, in word, right after the KWIC position, with KWIC inserted] ‘[Option D] ‘social social’ pressure’ yeah. [Line 7, in phrase, with KWIC inserted] ‘In fact, [Option D] ‘social’ scientists and historians?’ [P] Think it is. Hmm. [P] This one, [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] ‘Islamic anthropology and [Option D] ‘social’ analysis?’ Hmm. [P] Yeah. [P] [Line 2, in part, with KWIC inserted] ‘[IA] ‘other sexual drug use behavior’] and [Option D] ‘social’ and’ [P] Yeah. [P] I think I will choose the [Option D] ‘social’ one. It makes sense and also [P] because it it has ‘Personal’ [P] and ‘Emotion[al]’ [both from Line 5]. So, it should be in the same level of of this words.

R: Uh-huh.

Nina: And in the same like in [P] I dunno in the same environmental? Or something like the same group?

R: Uh-huh.

Nina: So, I will choose [Option D] ‘social’ yeah. [P] So, should I write? [chuckling]

R: Yes, yes. [chuckling] OK.

Nina: [Writing the word [Option D] ‘social’ on the answer slot] OK.

R: Any other key words that that really help you?

Nina: Well, it just helps me when you gave me the words [i.e., the options] and and, sure the fifth question [i.e., the concordance line] was was very helpful because it gave me the the background of or the group that the word should be [Note that Line 5 has a parallel structure of ‘Personal Behavior Difficulty’ and ‘Emotional Behavior Difficulty’, with the key forming part of the phrase ‘Social Behavior Difficulty’]. And, [P] I dunno. er just then it just making sense, reading in, putting the word er

R: Uh-huh.

Nina: Maybe maybe here also [the KWIC blank of Line 4 being underlined and the word ‘political’ being marked] because we have ‘political’ [from Line 4], so we have something related also to this group. And ‘demographic’ [from Line 2] [the word ‘demographic’ in Line 2 being marked] it’s we have we have like demographic, social, ethic and everything. [P] Also the ‘characteristics’ [from Line 2], so it should be [P] also in the same group. [P] So, I think that’s it. But I can also [P] [The options being pointed at] I can only do with with the words that you gave me. [laughing]

R: Yeah, yeah. No worries.

Nina: Yeah.

R: Alright. Anything else.

Nina: No. I don't think so.

ConCloze 7: Maya on Item 2 (Target word: *traumatic*)

Maya: Erm. One [Line 1, in part] 'and therefore it is his [KWIC position here] childhood experience ['that is to' not verbalized]' [Line 2, in part] 'comparative framework and not ['other' not verbalized] [KWIC position here] contexts' ['such as Rwanda or' not verbalized] [Line 3, in part] '[IA 'her'] rape but the enduring [KWIC position here] [P] effects' [the rest of Line 3 not verbalized] [Line 4] 'has been fully experienced [P] [KWIC position here] ['the' not verbalized] event moves from the present' [Line 5] 'around two points: first, that [KWIC position here] events have usually been those' [P] [Line 6, in part] 'was clear that the woman's something [i.e., KWIC position] experience during World War ['II' not verbalized]' erm [Line 7] 'of coping with and expressing [KWIC position here] issues and experiences. The individual' [After all the concordance lines had been read.] Ooh, OK. So, maybe this time's around [P] er it's got something to do with [E] feelings. [E]

Researcher: Uh-huh.

Maya: [P] What word would fit into feelings? [P] [Line 7 again immediately, in part, left-hand only] 'coping with and expressing [KWIC position]' ['of coping with and expressing' in Line 7 being underlined] [P]

R: Keep saying whatever you are thinking.

Maya: [P] Well, this point I've got mine [IA 'blank'] So, I'm trying to think of the words. [P] [Line 7 again immediately, in part] 'coping [P] ['with and expressing' not verbalized] [KWIC position here] issues and experience ['issues and experience' in Line 7 being underlined]' [Line 6, in word, right after the KWIC position] '[KWIC position here] experience' ['experience' in Line 6 also being underlined] [Line 5, in word, right after the KWIC position] '[KWIC position here] events' ['events' in Line 5 being underlined] [sigh] [P] What can relate to events? And experience? [Maya was now writing 'events & experience?' in the space on the right-hand side] er [P] Psst. [P] This is really hard. [P] I think, again, I might not actually have the word. [P]

R: Keep saying.

Maya: Erm. [P] That's it. That I really. [The pen is put down on the desk.] Something relates to experience and words. [P]

R: OK. Now I'm giving you choices. [The option sheet is placed before Maya.] [P] OK, tell me what whatever you are thinking.

Maya: [P] [Option A] 'anxious', [Option B] 'conditioned', [Option C] 'disagreeable', [Option D] 'traumatic' [P] [Line 1, in phrase, from right before the KWIC position, with KWIC inserted] '[IA 'it is his'] [Option D] 'traumatic' childhood experience' [P] no, [P] maybe not. [P] No, it is [Option D] 'traumatic' [P] It is [Option D] 'traumatic' because [P] erm [P] here you have negative feelings. [P] erm, so, [Option C] 'disagreeable' wouldn't be there because that's, to say that you have [KWIC filled out in discrete phrase, without any particular concordance line identifiable] '[Option C] 'disagreeable' feelings' does not make sense.

R: Uh-huh.

Maya: Erm, [P] then [Option B] 'conditioned' [P] er, no, the context [Line 3, in part, with KWIC inserted] 'her rape but the enduring [Option B] 'conditioned' effect' doesn't make sense in this sentence.

R: Uh-huh.

Maya: Erm, so, [Option A] 'anxious' [P] again, it's a good word, but it's not very specific in relation to [P] [her own summarized key words pointed at] events and experience, [Option D] 'traumatic'. Because it seems like in all of these sentences, [Option D] 'traumatic' seems to be highlighting that this is a certain event and this is a certain experience that they [P] have been [P] received. [P] so, [P] that's why I use [Option D] 'traumatic'.

R: OK. [P] Anything else you would like to add?

Maya: No.

ConCloze 7: Nina on Item 2 (Target word: *traumatic*)

Nina: [Line 1, in part] ‘and therefore it is his [KWIC position here] [P] childhood experience’ [P] [the word ‘experience’ being underlined] [Line 1 again immediately, in word, right-hand only] ‘experience’ [P] [Line 2, in part] ‘comparative framework and not other [KWIC position here] [P] [IA ‘contexts such as’]’ [the word ‘contexts’ being underlined] [Line 3] ‘her rape but the enduring [P] something [i.e., KWIC position] effects [IA] on her’ [Line 4] ‘has been fully experienced [P] the [KWIC position here] events move [‘from the’ not verbalized] present’ [Line 5] ‘around two points, first that [P] this events [P] [‘have’ not verbalized] usually been [E] those [E]’ [P] [Line 6] ‘was clear that the woman’s [KWIC position here] [P] experience during World War Two’ [P] [Line 7] ‘of coping with and expressing [KWIC blank skipped] issues and experiences. [P] The individual’ [P] [Item instructions] ‘All the lines above miss the same word. [IA ‘What should be that word?’]’ [P] ‘experience’ is also here [‘experienced’ in Line 4 being underlined] [P] and here [‘experience’ in Line 6 being underlined] and here [‘experiences’ in Line 7 being underlined]. That sounds good. [Chuckling] [P] same [P] and then [IA ‘enduring somewhere?’] hmm. [Line 6, in part] ‘was clear that the woman’s [KWIC position here] [P] experience’ No, I have no clue. [P] Other again, yeah. It’s like we have like the same words all these lines, but I can’t find the words that fit in all of them.

Researcher: Uh-huh.

Nina: No, no, just just see if you give me ones [Giggling] now?

R: And and what, do you have any general ideas about these lines?

Nina: Like [P] if if they talk er like the same thing? Or or, are you saying like if I get if I have any idea of the word that could be? Or something?

R: Or some summary. I dunno. Up to you. Any [P] impression you get?

Nina: No, no.

R: Now I’m giving you four choices [the option sheet placed before Nina].

Nina: Ok. So, [Option A] ‘anxious’, [Option B] ‘conditioned’, [Option C] ‘disagreeable’, [Option D] ‘traumatic’. [P] [NB: Nina now moved the option sheet up to right under the concordance, and accordingly over the question stem and answer line.] Hmm, [Option D] ‘traumatic’ sounds good. [P] [Line 1, in part, with KWIC inserted] ‘therefore it is his [P] [Option D] ‘traumatic’ childhood experience’, yeah. [P] [Line 2, in part, with KWIC inserted] ‘framework and not other [Option D] ‘traumatic’ contexts’ [P] maybe? [Line 3, in part, with KWIC inserted] ‘rape but the enduring [Option D] ‘traumatic’ effects’ [P] yeah. I think I will go with [Option D] ‘traumatic’. [P]

R: Uh-huh.

Nina: Because [P] it is [‘effects’ in Line 3 being underlined] an effect for sure for something that happened [P] before. And also an experience, like childhood [IA ‘a bad’] childhood [from Line 1], or the [IA ‘good’] no such thing, also. I think [Option D] ‘traumatic’ is a good word. [P] Yeah, just because it fits haha [laughing]. And it sounds good er like reading erm out loud and [P] it makes sense, all the lines after putting the word [Option D] ‘trau traumatic’ in this. [P] So, I think I will go with [Option D] ‘traumatic’. [P]

R: OK.

Nina: [Writing the word [Option D] ‘traumatic’ on the answer line] Yeah.

R: OK, any other thing you would like to say?

Nina: No, I don’t think so. No, I’ve said everything. [chuckling]

R: OK. Thank you so much.

References

- Ackerman, P.L., Beier, M.E. and Bowen, K.R., 2000. Explorations of crystallized intelligence: completion tests, cloze tests, and knowledge. *Learning and Individual Differences*, **12**(1), 105–21.
- Adolphs, S. and Durow, V., 2004. Social–cultural integration and the development of formulaic sequences. In: Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, 107–26.
- Aiken, L.R., 1979. Relationship between the item difficulty and discrimination indexes. *Educational and Psychological Measurement*, **39**, 821–4.
- Alderson, J.C., 2000. Technology in testing: The present and the future. *System*, **28**(4), 593–603.
- Alderson, J.C., 1996. Do corpora have a role in language assessment? In: Thomas, J. and Short, M. (eds.), *Using Corpora for Language Research: Studies in the Honor of Geoffrey Leech*. Essex: Longman, 248–59.
- Alderson, J.C., 1990. Testing reading comprehension skills (part two): Getting students to talk about taking a reading test (a pilot study). *Reading in a Foreign Language*, **7**(1), 465–503.
- Alderson, J.C., 1980. Native and nonnative speaker performance on cloze tests. *Language Learning*, **30**(1), 59–76.
- Alderson, J.C. and Kremmel, B., 2013. Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, **30**(4), 535–56.
- American Educational Research Association, 1999. *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Anderson, N.J., Bachman, L., Perkins, K. and Cohen, A., 1991. An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, **8**(1), 41–66.
- Anthony, L., 2014. *AntConc (Version 3.4.3) [Computer Software]*. Tokyo, Japan: Waseda University.
- Armstead, P.K., 2001. *What Factors Affect Cheating in Secondary School and Why?* Ph.D. dissertation, University of Plymouth.
- Aston, G., 2002. Getting one's teeth into a corpus. In: Tan, M. (ed.), *Corpus Studies in Language Education*. Bangkok: Institute for English Language Education Press, 131–43.
- Aston, G., Bernardini, S. and Stewart, D., 2004. *Corpora and Language Learners*. Amsterdam: John Benjamins.
- Aull, L., 2015. *First-Year University Writing: A Corpus-Based Study with Implications for Pedagogy*. London: Palgrave Macmillan.
- Bachman, L.F., 2000. Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, **17**(1), 1–42.
- Bachman, L.F., 1991. What does language testing have to offer? *TESOL Quarterly*, **25**(4), 671–704.
- Bachman, L.F., 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L.F. and Palmer, A.S., 2010. *Language Assessment in Practice*. Oxford: Oxford University Press.
- Bachman, L.F. and Palmer, A.S., 1996. *Language Testing in Practice*. Oxford: Oxford University Press.
- Bachman, L.F. and Palmer, A.S., 1989. The construct validation of self-ratings of communicative language ability. *Language Testing*, **6**(1), 14–29.
- Baghaei, P. and Amrahi, N., 2011. The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, **53**(2), 192–211.
- Barker, F., 2006. *Corpora and Language Assessment: Trends and Prospects*. 26. Cambridge, UK: University of Cambridge Local Examinations Syndicate.

- Barlow, M., 2004. Software for corpus access and analysis. In: Sinclair, J.M. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 205–21.
- Batty, A., 2012. Identifying dimensions of vocabulary knowledge in the Word Associates Test. *Vocabulary Learning & Instruction*, **1**(1), 70–7.
- Belcher, D.D., 2006. English for specific purposes: Teaching to perceived needs and imagined futures in worlds of work, study, and everyday life. *TESOL Quarterly*, **40**(1), 133–56.
- Bensoussan, M. and Laufer, B., 1984. Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, **7**(1), 15–31.
- BERA, 2011. *Ethical Guidelines for Educational Research*. London: British Educational Research Association.
- Bhatia, V.K., 2008. Genre analysis, ESP and professional practice. *English for Specific Purposes*, **27**, 161–74.
- Bhatia, V.K., 1993. *Analyzing genre: Language use in professional settings*. Oxon, UK: Routledge.
- BNC Consortium, 2005. *The BNC Sampler (XML Version)*. Oxford, UK: Oxford University Computing Services.
- Bowles, M.A., 2010. *The Think-Aloud Controversy in Second Language Research*. New York: Routledge.
- Bowles, M.A. and Leow, R.P., 2005. Reactivity and type of verbal report in SLA research methodology. *Studies in Second Language Acquisition*, **27**, 415–40.
- Boyd, J., 2011. *The Role of Digital Devices in Vocabulary Acquisition*. Research Notes. 44. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Boyle, A. and Booth, D., 2000. *The UCLES/CUP Learner Corpus*. 1. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Brown, J.D., 2000. What is construct validity? *Shiken: JALT Testing & Evaluation SIG Newsletter*, **4**(2), 8–12.
- Brown, J.D., 1996. *Testing in Language Programs*. NJ, USA: Prentice Hall Regents.
- Brown, J.D., 1980. Relative merits of four methods for scoring cloze tests. *The Modern Language Journal*, **64**(3), 311–7.
- Buck, G., 1991. The testing of listening comprehension: An introspective study. *Language Testing*, **8**(1), 67–91.
- Buja, A. and Eyuboglu, N., 1992. Remarks on parallel analysis. *Multivariate Behavioral Research*, **27**(4), 509–40. Available from: <http://www.math.pku.edu.cn/teachers/yaoy/reference/buja1992.pdf>.
- Butler, J., 1991. Cloze procedures and concordances: The advantages of discourse level authenticity in testing expectancy grammar. *System*, **19**(1–2), 29–38.
- Cambridge English Language Assessment, 2016a. *Cambridge English: Advanced (CAE)*. [Retrieved November 30, 2016]. Available from: <http://www.cambridgeenglish.org/exams/advanced/>.
- Cambridge English Language Assessment, 2016b. *Cambridge English: Advanced (CAE): Exam Format*. [Retrieved November 30, 2016]. Available from: <http://www.cambridgeenglish.org/exams/advanced/exam-format/>.
- Cambridge English Language Assessment, 2010a. *Cambridge English: Advanced (CAE): What's in the Exam?* UCLES. [Retrieved Dec 14, 2013]. Available from: <http://www.cambridgeenglish.org/exams-and-qualifications/advanced/whats-in-the-exam/>.
- Cambridge English Language Assessment, 2010b. *Find Free Resources: Cambridge English: Advanced (CAE): Use of English*. UCLES. [Retrieved Nov 15, 2013]. Available from: <http://www.cambridgeenglish.org/images/26800-cae-use-of-english-sample-test.pdf>.
- Capel, A., 2010. *Insights and Issues Arising from the English Profile Wordlists Project*. Research Notes. 41. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Chapelle, C.A., 2012. Validity argument for language assessment: The framework is simple.... *Language Testing*, **29**(1), 19–27.

- Chapelle, C.A. and Abraham, R.G., 1990. Cloze method: What difference does it make? *Language Testing*, 7(2), 121–46.
- Chapelle, C.A., Enright, M.K. and Jamieson, J., 2010. Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.
- Charles, M., 2007. Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, 6(4), 289–302.
- Choi, I., Kim, K.S. and Boo, J., 2003. Comparability of a paper-based language test and a computer-based language test. *Language Testing*, 20(3), 295–320.
- Cobb, T., ca. 2011. *Frequency Based Vocabulary Tests*. Montreal: Tom Cobb; Université du Québec à Montréal. Available from: <http://www.lextutor.ca/tests/>.
- Cobb, T., 2013. *Cloze Passage Builders*. Université du Québec à Montréal. 2013, Available from: <http://www.lextutor.ca/cloze/>.
- Cobb, T., 1997. Is there any measurable learning from hands-on concordancing? *System*, 25(3), 301–15.
- Cohen, A.D., 2012. Test-taking strategies and task design. In: Fulcher, G. and Davidson, F. (eds.), *The Routledge Handbook of Language Testing*. Oxon: Routledge, 262–77.
- Cohen, L., Manion, L. and Morrison, K., 2011. *Research Methods in Education*. 7th ed. London: Routledge.
- Cortina, J.M., 1993. What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78(1), 98–104.
- Coxhead, A., 2010. What can Corpora Tell Us about English for Academic Purposes? In: O'Keeffe, A. and McCarthy, M. (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon, UK: Routledge, 458–70.
- Coxhead, A., 2000. A New Academic Word List. *TESOL Quarterly*, 34(2), 213–38.
- Cronbach, L.J. and Meehl, P.E., 1955. Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4), 281–302.
- Crossley, S.A., Louwse, M.M., McCarthy, P.M. and McNamara, D.S., 2007. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1), 15–30.
- Crossley, S.A., Salsbury, T. and McNamara, D.S., 2014, in print. Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Language Testing*, 1–22.
- Crossley, S.A., Salsbury, T. and McNamara, D.S., 2011. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29(2), 243–63.
- Crystal, D., 1997, 2003. *English as a Global Language*. 2nd ed. Cambridge, UK: Cambridge University Press.
- Cummins, J., 2000. *Putting Language Proficiency in Its Place: Responding to Critiques of the Conversational/Academic Language Distinction*. Available from: <http://www.iteachilearn.com/cummins/converacademlangdisti.html>.
- D'Agostino, J., 2005. Measuring learning outcomes: Reliability and validity issues. In: Phye, G.D., Levin, J. and Robinson, D.H. (eds.), *Empirical Methods for Evaluating Educational Interventions*. Burlington, MA, USA: Academic Press, 113–46. Available from: <http://site.ebrary.com/lib/leicester/docDetail.action?docID=10138409&ppg=1>.
- Davies, M., 2008–. *The Corpus of Contemporary American English: 450 Million Words, 1990–present*. Available from: <http://corpus.byu.edu/coca/>.
- Dillon, D.R., 2013. Grounded Theory and qualitative research. In: Chapelle, C.A. (ed.), *The Encyclopedia of Applied Linguistics (Vol. 4)*. West Sussex, UK: Wiley-Blackwell, 2420–7.
- Dorans, N.J., Moses, T.P. and Eignor, D.R., 2010. *Principles and Practices of Test Score Equating*. ETS RR-10-29. New Jersey: Educational Testing Service.
- Dörnyei, Z. and Ryan, S., 2015. *The Psychology of the Language Learner Revisited*. New York: Taylor & Francis.

- Elliott, M., 2010. *The Expression of Affect in Spoken English*. Research Notes. 42. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Embretson, S.E., 2007. Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, **36**(8), 449–55.
- Ericsson, K.A. and Simon, H.A., 1993. *Protocol Analysis: Verbal Reports as Data*. Revised ed. ed. MA, USA: Massachusetts Institute of Technology.
- Flowerdew, J., 1996. Concordancing in language learning. In: Pennington, M.C. (ed.), *The Power of CALL*. Texas: Athelstan, 97–113.
- Flowerdew, L., 2012. *Corpora and Language Education*. Hampshire, UK: Palgrave Macmillan.
- Flowerdew, L., 2000. Using a genre-based framework to teach organizational structure in academic writing. *ELT Journal*, **54**(4), 369–78.
- Franklin, S.B., et al., 1995. Parallel analysis: A method for determining significant principal components. *Journal of Vegetation Science*, **6**(1), 99–106.
- Fulcher, G., 2003a. Interface design in computer-based language testing. *Language Testing*, **20**(4), 384–408.
- Fulcher, G., 2003b. *Testing Second Language Speaking*. Harlow, UK: Pearson Education.
- Fulcher, G. and Davidson, F., 2007. *Language Testing and Assessment*. Oxon, UK: Routledge.
- Fulcher, G. and Svalberg, A., 2013. Limited aspects of reality: Frames of reference in language assessment. *International Journal of English Studies*, **13**(2): 1–19.
- Gabrenya, W.K., Jr., 2003. *Reliability*. 1.0th ed. Florida Institute of Technology. Available from: <http://my.fit.edu/~gabrenya/IntroMethods/eBook/reliability.pdf>.
- Galaczi, E. and Khalifa, H., 2009. *Cambridge ESOL's CEFR DVD of Speaking Performances: What's the Story?* Research Notes. 37. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Gardner, D. and Davies, M., 2014. A new academic vocabulary list. *Applied Linguistics*, **35**(3), 305–27.
- Gaskell, D. and Cobb, T., 2004. Can Learners use Concordance Feedback for Writing Errors? *System*, **32**, 301–19.
- Gilquin, G. and Granger, S., 2010. How can data-driven learning be used in language teaching? In: O'Keeffe, A. and McCarthy, M. (eds.), *The Routledge handbook of corpus linguistics*. Abington, UK: Routledge, 359–70.
- Graddol, D., 1997. *The Future of English? A Guide to Forecasting the Popularity of the English Language in the 21st Century*. 2000th ed. London: The British Council.
- Greaves, C. and Warren, M., 2010. What can a corpus tell us about multi-word units? In: O'Keeffe, A. and McCarthy, M. (eds.), *The Routledge handbook of corpus linguistics*. London: Routledge, 212–26.
- Green, A., 2008. *English Profile: Functional Progression in Materials for ELT*. Research Notes. 33. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Gyllstad, H., 2009. Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In: Barfield, A.W. and Gyllstad, H. (eds.), *Researching Collocations in Another Language: Multiple Interpretations*. Hampshire, UK: Palgrave Macmillan, 153–70.
- Gyllstad, H., 2007. *Testing English Collocations: Developing Receptive Tests for use with Advanced Swedish Learners*. Ph.D. dissertation, Lund University.
- Haladyna, T.M., 2004. *Developing and Validating Multiple-Choice Test Items*. 3rd ed. NJ, USA: Lawrence Erlbaum.
- Hanania, E. and Shikhani, M., 1986. Interrelationships among three tests of language proficiency: Standardized ESL, cloze, and writing. *TESOL Quarterly*, **20**(1), 97–109.
- Hancioğlu, N., Neufeld, S. and Eldridge, J., 2008. Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes*, **27**, 459–79.
- Hargreaves, P., 2000. Collocation and testing. In: Lewis, M. (ed.), *Teaching collocation: Further developments in the lexical approach*. Massachusetts, USA: Thomson, 205–23.

- Hawkey, R., 2001. *Towards a Common Scale to Describe L2 Writing Performance*. 5. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Henning, G.H., 1987. *A Guide to Language Testing: Development, Evaluation, Research*. USA: Newbury House.
- Hoey, M., 2013. Lexical priming. In: Chapelle, C.A. (ed.), *The Encyclopedia of Applied Linguistics* (Vol. 6). West Sussex, UK: Wiley-Blackwell, 3342–7.
- Hoey, M., 2005. *Lexical Priming: A New Theory of Words and Language*. Oxon, UK: Routledge.
- Hoey, M., 2000. A world beyond collocation: New perspectives on vocabulary teaching. In: Lewis, M. (ed.), *Teaching collocation: Further developments in the lexical approach*. Massachusetts, USA: Thomson, 224–43.
- Hoffman, L., Yang, X., Bovaird, J.A. and Embretson, S.E., 2006. Measuring attentional ability in older adults: Development and psychometric evaluation of DriverScan. *Educational and Psychological Measurement*, **66**(6), 984–1000.
- Hoshino, A., 2009. *Automatic Question Generation for Language Testing and its Evaluation Criteria*. Ph.D. dissertation, Graduate School of Interdisciplinary Information Studies, University of Tokyo.
- Housen, A., Kuiken, F. and Vedder, I., 2012. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins.
- Hughes, G., 2006. *The Effect of Editing on Language used in FCE Reading Texts: A Case Study*. 26. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Hupé, J., 2015. Statistical inferences under the null hypothesis: Common mistakes and pitfalls in neuroimaging studies. *Frontiers in Neuroscience*, **9**(18).
- Huynh, H. and Meyer, P., 2010. Use of Robust Z in detecting unstable items in Item Response Theory models. *Practical Assessment, Research & Evaluation*, **15**(2).
- Instructional Assessment Resources, 2011. *Item Analysis*. Available from: <https://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php>.
- Jackson, T.R., et al., 2002. Validation of authentic performance assessment: A process suited for Rasch modeling. *American Journal of Pharmaceutical Education*, **66**(3), 233–42.
- Jasper, F., 2010. Applied dimensionality and test structure assessment with the START-M mathematics test. *The International Journal of Educational and Psychological Assessment*, **6**(1), 104–25.
- Jenkins, J., 2003. *World Englishes: A Resource Book for Students*. London: Routledge.
- John, O.P. and Benet-Martínez, V., 2000. Measurement: Reliability, construct validation, and scale construction. In: Reis, H.T. and Judd, C.M. (eds.), *Handbook of Research Methods in Social and Personality Psychology*. New York: Cambridge University Press, 339–69.
- Johnstone, C.J., Bottsford-Miller, N.A. and Thompson, S.J., 2006. *Using the Think Aloud Method (Cognitive Labs) to Evaluate Test Design for Students with Disabilities and English Language Learners*. Technical report. 44. Minneapolis, Minnesota: University of Minnesota, National Center on Educational Outcomes.
- Kane, M.T., 2012a. All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives*, **10**(1–2), 66–70.
- Kane, M.T., 2012b. Validating Score Interpretations and Uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, **29**(1), 3–17.
- Kane, M.T., 2006. Validation. In: Brennan, R.L. (ed.), *Educational measurement*. 4th ed. Connecticut, US: American Council on Education; Praeger, 17–64.
- Kane, M.T., 1992. An argument-based approach to validity. *Psychological Bulletin*, **112**(3), 527–35.
- Kane, M.T., 1990. *An Argument-Based Approach to Validation*. ACT-RR-90-13. Iowa City: The American College Testing Program.
- Kane, M.T., Crooks, T. and Cohen, A., 1999. Validating measures of performance. *Educational Measurement: Issues and Practice*, **18**(2), 5–17.

- Kenyon, D.M. and MacGregor, D., 2012. Pre-operational testing. In: Fulcher, G. and Davidson, F. (eds.), *The Routledge handbook of language testing*. Oxon: Routledge, 295–306.
- Kilgariff, A. 2009. Corpora in the classroom without scaring the students. *Proceedings of the 18th international symposium on English teaching, Taipei*. Anonymous Available from: <http://www.kilgariff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc>.
- Kim, N., 2016. *Test Preparation and Identity: Cases of Korean Test Takers Preparing Tests of English Speaking Proficiency*. Ph.D. dissertation, University of Leicester.
- King, J., 2013. *Silence in the Second Language Classroom*. New York: Palgrave Macmillan.
- Kline, P., 1991. *Intelligence: The Psychometric View*. London: Routledge.
- Kongsuwannakul, K. 2015a. Prototyping a concordance-based cloze test: Preliminary results. *Proceedings of the 7th International Conference on Humanities and Social Sciences: ASEAN 2015: Challenges and Opportunities*. The Faculty of Liberal Arts, Prince of Songkla University, ed. Songkhla, Thailand. Available from: <http://fs.libarts.psu.ac.th/research/conference/proceedings-7/1/1.7-Prototyping%20a%20Concordance-based%20Cloze%20Test.pdf>.
- Kongsuwannakul, K., 2015b. Six techniques for creating variety in the concordance-based cloze item type. *International Journal of Assessment and Evaluation*, **22**(2), 23–35.
- Kongsuwannakul, K., 2014a. Language processes of the concordance-based cloze item type: Bridging a theoretical gap between language testing and second language acquisition. *The International Journal of Communication and Linguistic Studies*, **12**(1), 17–27.
- Kongsuwannakul, K., 2014b. Theoretical considerations of applications and implications of concordance-based cloze tests. *Digital Scholarship in the Humanities*, **30**(4), 541–58.
- Kuusela, H. and Paul, P., 2000. A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology*, **113**(3), 387–404.
- Lange, D.L. and Clausing, G., 1981. An examination of two methods of generating and scoring CLOZE tests with students of German on three levels. *The Modern Language Journal*, **65**(3), 254–61.
- Laufer, B., and Nation, P., 1999. A vocabulary-size test of controlled productive ability. *Language Testing*, **16**(1): 33–51.
- Lee, H.S. and Winke, P., 2013. The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, **30**(1), 99–123.
- Lewis, Michael, 2000. *Teaching Collocation: Further Developments in the Lexical Approach*. London: Thomson Heile.
- Linacre, J.M., ca. 2012. *Misfit Diagnosis: Infit Outfit Mean-Square Standardized: Winsteps Help*. [Retrieved July 11, 2013]. Available from: <http://www.winsteps.com/winman/diagnosingmisfit.htm>.
- Linacre, J.M., 2015a. *Equating and Linking Tests: Winsteps Help*. [Retrieved August 20, 2015]. Available from: <http://www.winsteps.com/winman/equating.htm>.
- Linacre, J.M., 2015b. *User-Friendly Rescaling: Zero Point and Unit*. [Retrieved August 20, 2015]. Available from: <http://www.winsteps.com/winman/rescaling.htm>.
- Linacre, J.M., 2012. Some Question on Equating and Linking Method. *Old Rasch Forum - Rasch on the Run: 2012*: 65. [Retrieved August 20, 2014]. Available from: <http://www.rasch.org/forum2012.htm>.
- Linacre, J.M., 2004. Test validity and Rasch measurement: Construct, content, etc. *Rasch Measurement Transactions*, **18**(1) 970–1.
- Linacre, J.M., 2002. Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, **3**(1) 85–106.
- Linacre, J.M., 1994. Dichotomous infit and outfit mean-square fit statistics. *Rasch Measurement Transactions*, **8**(2), 360.
- Loevinger, J., 1957. Objective tests as instruments of psychological theory. *Psychological Reports*, **3**, 635–94.
- Longman Dictionary of Contemporary English, 2015. Pearson. Available from: <http://www.ldoceonline.com/>.

- Luoma, S. and Tarnanen, M., 2003. Creating a self-rating instrument for second language writing: From idea to implementation. *Language Testing*, **20**(4), 440–465.
- Mason, M., 2010. Sample size and saturation in PhD studies using qualitative interviews. *Forum Qualitative Sozialforschung*, **11**(3), Article 8.
- McCarthy, M. and Carter, R., 1997. Written and spoken vocabulary. In: Schmitt, N. and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 20–39.
- McCarthy, M. and O'Keeffe, A., 2010. Historical perspective: What are corpora and how have they evolved? In: O'Keeffe, A. and McCarthy, M. (eds.), *The Routledge Handbook of Corpus Linguistics*. Abington, UK: Routledge, 3–13.
- McNamara, T., 2006. Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly: An International Journal*, **3**(1), 31–51.
- McNamara, T., 1996. *Measuring Second Language Performance*. Essex, UK: Longman.
- Meara, P.M., 2009. *Connected Words: Word Associations and Second Language Vocabulary Acquisition*. The Netherlands: John Benjamins.
- Messick, S., 1996. *Validity and Washback in Language Testing*. RR-96-17. New Jersey: Educational Testing Service.
- Messick, S., 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**(9), 741–9.
- Messick, S., 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, **23**(2), 13–23.
- Messick, S., 1993. *Foundations of Validity: Meaning and Consequences in Psychological Assessment*. RR-93-51. New Jersey: Educational Testing Service.
- Messick, S., 1989. Validity. In: Linn, R.L. (ed.), *Educational Measurement*. 3rd ed. New York: American Council on Education; Collier Macmillan, 13–103.
- Messick, S., 1988. The once and future issues of validity: Assessing the meaning and consequences of measurement. In: Wainer, H. and Braun, H.I. (eds.), *Test Validity*. New Jersey: Lawrence Erlbaum Associates, 33–45.
- Mislevy, R.J. and Riconscente, M.M., 2005. *Evidence-Centered Assessment Design: Layers, Structures, and Terminology*. 9. California: SRI International.
- Mislevy, R.J., Steinberg, L.S. and Almond, R.G., 1999. *Evidence-Centered Assessment Design*. New Jersey: Educational Testing Service.
- Mislevy, R.J., 2007. Validity by design. *Educational Researcher*, **36**(8), 463–9.
- Morgan, H. 1992. An analysis of Gardner's theory of multiple intelligence. *The Annual Meeting of the Eastern Educational Research Association*. Georgia, US: ERIC Document Reproduction Service.
- Moss, P.A., 2007. Reconstructing validity. *Educational Researcher*, **36**(8), 470–6.
- Moss, P.A., 1994. Can there be validity without reliability? *Educational Researcher*, **23**(2), 5–12.
- Mowbray, C.T., Holter, M.C., Teague, G.B. and Bybee, D., 2003. Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, **24**(3), 315–40.
- Nation, I.S.P., 2001. *Learning Vocabulary in another Language*. Cambridge, UK: Cambridge University Press.
- Nation, P. and Waring, R., 1997. Vocabulary size, text coverage and word lists. In: Schmitt, N. and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition, Pedagogy*. Cambridge: Cambridge University Press, 6–19.
- Neff-van Aertselaer, J., 2008. *Arguing in English and Spanish: A Corpus Study of Stance*. Research Notes. 33. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Newman, J.A., 2016. *A corpus-based comparison of the Academic Word List and the Academic Vocabulary List*. Master's thesis, Brigham Young University.

- Nielsen, J., 1994. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, **41**(3), 385–97.
- Nissan, S. and Schedl, M., 2012. Prototyping new item types. In: Fulcher, G. and Davidson, F. (eds.), *The Routledge Handbook of Language Testing*. Oxon: Routledge, 281–94.
- Nunan, D. 1991. Second language proficiency assessment and program evaluation. In: Anivan, S. (ed.), *Issues in Language Programme Evaluation in the 1990's*. Anthology Series 27, 46–59. ERIC Document No. 367152.
- O'Dell, F., 1997. Incorporating vocabulary into the syllabus. In: Schmitt, N. and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, 258–78.
- O'Loughlin, K., 2011. The interpretation and use of proficiency test scores in university selection: How valid and ethical are they?, *Language Assessment Quarterly*, **8**(2), 146–60.
- Ohnmacht, F.W., Weaver, W.W. and Kohler, E.T., 1970. Cloze and closure: A factorial study. *Journal of Psychology*, **74**(2), 205–17.
- Oller, J.W., Jr., 1979. *Language Tests at School: A Pragmatic Approach*. London: Longman Group.
- Oller, J.W., Jr. and Conrad, C.A., 1971. The cloze technique and ESL proficiency. *Language Learning*, **21**(2), 183–95.
- Pace-Sigge, M., 2013. The concept of lexical priming in the context of language use. *ICAME Journal*, **37**, 149–73.
- Papp, S., 2007. Inductive learning and self-correction with the use of learner and reference corpora. In: Hidalgo, E., Quereda, L. and Santana, J. (eds.), *Corpora in the Foreign Language Classroom: Selected Papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6), University of Granada, Spain, 4-7 July, 2004*. Amsterdam: Rodopi, 207–20.
- Paquot, M., 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.
- Park, K., 2014. Corpora and language assessment: The state of the art. *Language Assessment Quarterly*, **11**(1), 27–44.
- Parshall, C.G., Spray, J.A., Kalohn, J.C. and Davey, T., 2002. *Practical Considerations in Computer-Based Testing*. New York: Springer.
- Partington, A., 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
- Petre, M. and Rugg, G., 2010. *The Unwritten Rules of PhD Research*. Second ed. Berkshire, UK: Open University Press.
- Poole, R., 2011. *Concordance-based Glosses for Facilitating Semantization and Enhancing Productive Knowledge of Academic Vocabulary*. Master's thesis, University of Alabama.
- Porter-Szucs, I. and Jameel, U., 2014. *Nativelike Formulaic Sequences in Office Hours: Validating a Speaking Test for International Teaching Assistants*. Research Notes. 55. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Powers, D.E., 2010. The case for a comprehensive, four-skills assessment of English-language proficiency. *R&D Connections*, **14**.
- Pressley, M. and Afflerbach, P., 1995. *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading*. New Jersey, USA: Lawrence Erlbaum.
- Proudfoot, S., 2010. *A Corpus-Led Exploration of Lexical Verb use in Main Suite Writing Papers*. Research Notes. 41. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Purpura, J.E., 2004. *Assessing Grammar*. Cambridge: Cambridge University Press.
- Purpura, J.E., 1999. *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach*. Cambridge, UK: Cambridge University Press.
- Read, J., 2012. Piloting vocabulary tests. In: Fulcher, G. and Davidson, F. (eds.), *The Routledge Handbook of Language Testing*. Oxon: Routledge, 307–20.

- Read, J., 2007. Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies*, **7**(2), 105–25.
- Read, J., 2005. *Applying Lexical Statistics to the IELTS Speaking Test*. 20. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Read, J., 2000. *Assessing Vocabulary*. Cambridge, UK: Cambridge University Press.
- Read, J., 1998. Validating a test to measure depth of vocabulary knowledge. In: *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach*. New Jersey: Lawrence Erlbaum Associates, 41–60.
- Read, J., 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing*, **10**(3), 355–71.
- Read, J. and Chapelle, C.A., 2001. A framework for second language vocabulary assessment. *Language Testing*, **18**(1), 1–32.
- Richards, B., Malvern, D. and Graham, S., 2008. Word frequency and trends in the development of French vocabulary in lower-intermediate students during Year 12 in English schools. *The Language Learning Journal*, **36**(2), 199–213.
- Rigney, S.L., Wiley, D.E. and Kopriva, R.J., 2008. The past as preparation: Measurement, public policy and implications for access. In: Kopriva, R.J. (ed.), *Improving Testing for English Language Learners*. New York: Routledge, 37–63.
- Rodriguez, M.C., 2005. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, **24**(2), 3–13.
- Römer, U., 2011. Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, **31**, 205–25.
- Rose, D., 2008. *Vocabulary Use in the FCE Listening Test*. Research Notes. 32. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Ross, S.J., 2012. Claims, evidence, and inference in performance assessment. In: Fulcher, G. and Davidson, F. (eds.), *The Routledge Handbook of Language Testing*. Oxon: Routledge, 223–33.
- Ruiz-Garrido, M.F., Palmer-Silveira, J.C. and Fortanet-Gómez, I., 2010. Current trends in English for Professional and Academic Purposes. In: Ruiz-Garrido, M.F., Palmer-Silveira, J.C. and Fortanet-Gómez, I. (eds.), *English for Professional and Academic Purposes*. Amsterdam: Rodopi, 1–8.
- Saumure, K. and Given, L.M., 2008. Data saturation. In: Given, L.M. (ed.), *The Sage Encyclopedia of Qualitative Research Methods*. California: Sage, 195–6.
- Schacter, D.L., 1999. The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, **54**(3), 182–203.
- Schmitt, N., 2013. *Formulaic Language in Academic Study*. [PowerPoint presentation]. TESOL 2013 conference: [Retrieved March 23, 2016]. Available from: www.norbertschmitt.co.uk/uploads/27_514b291ddcf0a981038115.ppt.
- Schmitt, N., 2010. *Researching Vocabulary a Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.
- Schmitt, N., 2005–6. Formulaic language: Fixed and varied. *Estudios de Lingüística Inglesa Aplicada*, **6**, 13–39.
- Schmitt, N., 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.
- Schmitt, N., 1999. The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing*, **16**(2), 189–216.
- Schmitt, N. and Carter, R., 2004. Formulaic sequences in action: An introduction. In: Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, 1–22.
- Schmitt, N., Schmitt, D. and Clapham, C., 2001. Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test. *Language Testing*, **18**(1), 55–88.

- Schmitt, N., Dörnyei, Z., Adolphs, S. and Durow, V., 2004a. Knowledge and acquisition of formulaic sequences: A longitudinal study. In: Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, 54–86.
- Schmitt, N., Grandage, S. and Adolphs, S., 2004b. Are corpus-derived recurrent clusters psycholinguistically valid? In: Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, 126–51.
- Schmitt, N. and Meara, P.M., 1997. Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, **20**, 17–36.
- Schmitt, N., Ng, J.W.C. and Garras, J., 2011. The Word Associates Format: Validation evidence. *Language Testing*, **28**(1), 105–26.
- Schneider, Edgar W., 1997. *Varieties of English Around the World*. Philadelphia, US: John Benjamins.
- Schneiderman, W., 1980. A personality dimension of consistency versus variability without the use of self-reports or ratings. *Journal of Personality and Social Psychology*, **19**(1), 158–64.
- Schoonen, R. and Verhallen, M., 2008. The assessment of deep word knowledge in young first and second language learners. *Language Testing*, **25**(2), 211–36.
- Scott, M., 2012. *WordSmith Tools (Version 6)*. Stroud: Lexical Analysis Software.
- Shaw, S.D. and Weir, C.J., 2007. *Examining Writing: Research and Practice in Assessing Second Language Writing*. Cambridge: Cambridge University Press.
- Shiffrin, R.M. and Nosofsky, R.M., 1994. Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, **101**(2), 357–61.
- Shin, S., 2009. How to treat omitted responses in Rasch model-based equating. *Practical Assessment, Research & Evaluation*, **14**(1).
- Simpson-Vlach, R. and Ellis, N.C., 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, **31**(4), 487–512.
- Sinclair, John M., 2004a. *How to Use Corpora in Language Teaching*. Sinclair, J.M. (ed.), Amsterdam: John Benjamins.
- Sinclair, J.M., 2004b. New evidence, new priorities, new attitudes. In: Sinclair, J.M. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 271–99.
- Sinclair, J.M., 2003. *Reading Concordances: An Introduction*. London: Pearson Longman.
- Sinclair, J.M., 1991. *Corpus, Concordance, Collocation*. Oxford, UK: Oxford University Press.
- Sinclair, J.M. and Renouf, A., 1988. A lexical syllabus for language learning. In: Carter, R. and McCarthy, M. (eds.), *Vocabulary and Language Teaching*. London: Longman, 140–59.
- Spöttl, C. and McCarthy, M., 2004. Comparing knowledge of formulaic sequences across L1, L2, L3, and L4. In: Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, 191–219.
- Steele, C.M. and Aronson, J., 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, **69**(5), 797–811.
- Stevens, V., 1991a. Classroom concordancing: Vocabulary materials derived from relevant, authentic text. *English for Specific Purposes*, **10**, 35–46.
- Stevens, V., 1991b. Concordance-based vocabulary exercises: A viable alternative to gap-fillers. In: Johns, T. and King, P. (eds.), *English Language Research Journal*, Vol. 4. Birmingham: University of Birmingham, 47–61.
- Stubbs, M., 2002. *Words and Phrases*. Oxford, UK: Blackwell.
- Szudarski, P., 2013. Lexical syllabus. In: Chappelle, C.A. (ed.), *The Encyclopedia of Applied Linguistics* (Vol. 6). West Sussex, UK: Wiley-Blackwell, 3356–60.
- Taylor, L. and Barker, F., 2008. Using corpora for language assessment. In: Shohamy, E. and Hornberger, N.H. (eds.), *Encyclopedia of Language and Education*. 2nd ed. New York: Springer Science & Business Media, 241–54.

- Teubert, W. and Čermáková, A., 2004. *Corpus Linguistics: A Short Introduction*. London: Continuum.
- Thurstun, J. and Candlin, C.N., 1998. Concordancing and the teaching of the vocabulary of Academic English. *English for Specific Purposes*, **17**(3), 267–80.
- Tognini-Bonelli, E., 2001. *Corpus Linguistics at Work*. John Benjamins Publishing Available from: <http://www.jbe-platform.com/content/books/9789027285447>.
- Tribble, C., 2013. Concordancing. In: Chapelle, C.A. (ed.), *The Encyclopedia of Applied Linguistics* (Vol. 2). West Sussex, UK: Wiley-Blackwell, 871–9.
- Tribble, C. 1997. Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching. *Practical Applications in Language Corpora (PALC)*. Melia, J. and Lewandowska-Tomaszczyk, B., eds. Lodz, Poland.
- Tribble, C. and Jones, G., 1990. *Concordances in the Classroom: A Resource Book for Teachers*. Harlow: Longman.
- UK Copyright Service., 2009. *Fact Sheet P-27: Using the Works of Others*. Available from: http://www.copyrightservice.co.uk/copyright/p27_work_of_others#fair_dealing.
- UK Copyright Service., 2004. *Copyright Law Fact Sheet P-09: Understanding Fair Use*. Available from: http://www.copyrightservice.co.uk/copyright/p09_fair_use.
- Vollmer, H.J., 1983. The structure of foreign language competence. In: Hughes, A. and Porter, D. (eds.), *Current Developments in Language Testing*. London: Academic Press, 3–34.
- West, M., 1953. *A General Service List of English Words*. London: Longman.
- White, P., 2009. *Developing Research Questions: A Guide for Social Scientists*. Hampshire: Palgrave Macmillan.
- Wilhelm, K.H., 1996. Combined Assessment Model for EAP Writing Workshop: Portfolio Decision-Making, Criterion-Referenced Grading, and Contract Negotiation. *TESL Canada Journal*, **14**(1), 21–33.
- Willis, J., 2011. Concordances in the classroom without a computer: Assembling and exploiting concordances of common words. In: *Materials Development in Language Teaching*. Cambridge: Cambridge University Press, 51–77. Available from: <http://www.cambridge.org/servlet/file/store7/item7138221/version1/Tomlinson%20Willis%20Chap%203.pdf>
- Wood, D., 2016. *Fundamentals of Formulaic Language: An Introduction*. London: Bloomsbury Academic.
- Woolard, G., 2000. Collocation – encouraging learner independence. In: Lewis, M. (ed.), *Teaching Collocation: Further Developments in the Lexical Approach*. Massachusetts, USA: Thomson, 28–46.
- Wray, A., 2002. *Formulaic Language and the Lexicon*. Cambridge, UK: Cambridge University Press.
- Wright, A., 2008. *A Corpus-Informed Study of Specificity in Financial English: The Case of ICFE Reading*. Research Notes. 31. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Wright, B. and Linacre, J.M., 1994. Reasonable mean-square fit values. *Rasch Measurement Transactions*, **8**(3), 370.
- Yang, B., Watkins, K.E. and Marsick, V.J., 2004. The construct of the learning organization: Dimensions, measurement, and validation. *Human Resource Development Quarterly*, **15**(1), 31–55.
- Zinkhan, G.M. and Martin, C.R., Jr., 1983. Two copy testing techniques: The cloze procedure and the cognitive complexity test. *Journal of Business Research*, **11**(2), 217–27.