Language learning in complex virtual worlds: Effects of modality and task complexity on oral performance between virtual world and face-to-face tasks.

Thesis submitted for the degree of

Doctor of Philosophy

in

Education Research

at the University of Leicester

by

James York

School of Education

University of Leicester

2019

Language learning in complex virtual worlds: Effects of modality and task complexity on oral performance between virtual world and face-to-face tasks.

by James York

ABSTRACT

Virtual worlds have been identified as a potentially beneficial domain for language learning due to various cognitive and affective affordances such as immersive content, access to native speakers, and motivating properties. However, research on computer-mediated communication (CMC) has largely ignored the use of virtual worlds as a possible domain for communication. Additionally, the game-based language teaching (GBLT) sub-field of CALL has focused too narrowly on specific virtual world affordances, overlooking how communicating in such complex domains may affect learner output, particularly in comparison with face-to-face communication. Thus, the main aim of this study is to explore the potential differences in learner oral performance as they conduct tasks via two oral modalities: within a virtual world and face-to-face.

Twenty participants (10 dyads) conducted six dialogic tasks, organised by modality into three task-pairs. Quantitative data was collected via transcribing audio recordings of all sessions. The data were analysed in terms of learners' output complexity, accuracy and fluency using appropriate measures for each. Post-task questionnaires were employed to gauge perceptions of task difficulty, and therefore validate the researcher's presumptions of task complexity. This data was also used to provide insight into findings from the quantitative data.

Results suggest that virtual world tasks may hinder output fluency. However, complexity and accuracy were not significantly affected by mode. Instead, task complexity and type had a more considerable influence on these constructs. Lexical density was higher when conducting virtual world tasks, and, regardless of the increased cognitive demands posed by the virtual world, participants preferred to undertake tasks in this domain. Implications are provided regarding virtual world task design and the cognitive and affective affordances of virtual worlds for language learning, specifically for classroom contexts. Finally, the limitations of this study inform avenues for future research.

ACKNOWLEDGEMENTS

I want to thank my supervisor Pamela Rogerson-Revell for sticking with me through this arduous journey and keeping me on the right path. Your comments and guidance were invaluable in helping me achieve this monumental task. Gratitude is also extended to Jim Askham, who helped guide me into a specific direction by asking critical questions during my annual review.

Secondly, I would like to thank all of the participants in this study from Tokyo Denki University. Without their consent to collect data, this study would not exist. Similarly, I would like to thank Tero Korpela for helping me purchase, set up, and manage the server on which the virtual world exists and Kai Fukushima for inspiring me in regard to content creation.

My warm regards to Jonathan deHaan, Peter Hourdequin, Benjamin Thanyawatpokin and other members of the Japan Game Lab for being interested in my research and providing critical comments on results.

Finally, I would like to thank my family. First, my wife and children for allowing me to pursue this research at the expense of some quality family time, and, secondly, my mother and father for always allowing me to pursue my interests regardless of the strange and unpredictable decisions that I seem to make.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

ACMC	Asynchronous computer-mediated communication
AT	Activity theory
CALL	Computer-assisted language learning
CH	Cognition hypothesis
CLT	Communicative language teaching
CMC	Computer-mediated communication
COTS	Commercial off the shelf (game)
DGBLL	Digital game based language learning
EFL	English as a second language
FTF	Face to face
ICT	Information communication technology
ISLA	Instructed SLA
L1	First language
L2	Second language
LACM	Limited attentional capacity model
MMO	Shortened form of MMORPG
MMORPG	Massively multiplayer online role-playing game
MOO	MUD, object-oriented
MTLD	Measure of textual lexical diversity
MUD	Multi-user dungeon
MUVE	Multi-user virtual environment
PPP	Present, practice, perform
RPG	Role-playing game
SCMC	Synchronous computer mediated communication
SIE	Synthetic immersive environment
SLA	Second language acquisition
SPSS	Statistical Package for the Social Sciences
TBLT	Task-based language teaching
TENOR	Teaching English for no obvious reason
TESOL	Teaching English as a second language
VOIP	Voice over internet protocol
VW	Virtual world
WoW	World of Warcraft

1 INTRODUCTION

1.1 BACKGROUND TO THE STUDY

Technology is having a profound effect on education, with an increasing amount of information and communication technology (ICT) usage in practically every field. Examples of the range and scope of technologies used in classrooms around the world have been documented in an equally growing number of books on the subject (Balacheff, Ludvigsen, de Jong, Lazonder, & Barnes, 2009; Berger & Trexler, 2010). Further, in second language (L2) learning contexts, there is a growing number of research papers and books exploring pedagogical experimentation with technology (Thomas, 2009; Thomas & Reinders, 2010, González & Ortega, 2014; Chapelle & Sauro, 2017; Sykes, 2018). The current study explores the potential of a particular technology which has the potential to transform how we currently learn and teach languages (Johnson, Smith, Willis, Levine, & Haywood, 2011). This technology is known by several names such as 3D multi-user environments, synthetic immersive environments, social worlds, digital games, or as I shall call them henceforth: virtual worlds (VWs).

Computer assisted language learning (CALL) literature on the use of digital games, and VWs has increased dramatically in the last decade. From 2000-2004, only .81% of articles in the Social Sciences Citation Index were written about the use of such games, and this figure rose to 3.82% between 2005 and 2009 (see Lin, 2015). Chapelle (2001) also writes that in the 21st century, technology-mediated tasks are of great importance for researchers and teachers interested in SLA. One reason for this was mentioned above: Digital, network-connected technology is affordable enough that it is becoming pervasive in all facets of our professional and private lives, transforming the way we communicate. Adoption in educational contexts is also rapidly increasing where the argument for its adoption is based on the belief that technology will improve the efficiency and effectiveness of education (George & Sanders, 2017). There are however concerns that the technology alone may not improve learning (OECD, 2015), promoting a call for empirical research on how technology may be successfully implemented in classrooms. Finally, the application of new technologies

in language learning contexts is being explored in edited volumes such as Gonzalez-Lloret and Ortega (2014). This volume is explicitly concerned with a technology-mediated approach to language teaching, representing a conceptual shift from technology as a tool or tutor, to technology as a method for successful language acquisition. A similar conceptualisation of technology and its role in language teaching has been outlined in Reinhardt and Thorne (2016) who explore the genesis of CALL and the evolution of technology from tool, to tutor, ecology and finally the method of instruction.

The current study aims to investigate the effects of performing interactive tasks within a VW on learners' oral task performance. The main rationale being that virtual worlds represent a particular point on the continuum of computer-mediated communication technologies, which have received little attention to date. Additionally, there is a particular dearth in research which explores both task-based instruction and learners' oral performance in such domains. A preference for text-based communication appears much more frequently in the CALL literature (for example, see Lin, 2014; Zeigler, 2016a, 2016b).

Participants in this study completed several tasks designed to utilise the affordances of both face-to-face (FTF) and VW modes of communication. Regarding these modalities, the first mode is that found in traditional classroom contexts: face-to-face communication. The second is categorised as synchronous computer-mediated communication (SCMC). Therefore, this study hopes to generate knowledge regarding the benefits or hindrances of using virtual environments in classroom contexts by comparing learners' oral task performance as they complete three sets of VW and FTF tasks. The tasks were operationalised to be of low, medium and high task complexity, in order to explore the effect of modality and task complexity on learner performances.

Additionally, tasks were designed to be as similar as possible within a task pair. However, the affordances of both modes of communication were considered during the design phase, meaning that cognitive task complexity differences between tasks in a task pair were unavoidable. In order to validate task complexity predictions, a cognitive load questionnaire was utilised to measure learners' perceptions of cognitive load for each task.

1.2 WHY ADVOCATE FOR VIRTUAL WORLDS?

From a task-based language learning and teaching (TBLT) perspective, which

typically promotes the use of real-world, authentic tasks in the target language, the activities that students undertake in virtual worlds may be considered authentic language-learning tasks. This is due to the rich, meaningful discourses that occur on several levels when players interact within such environments (González-Lloret, 2015b). Virtual worlds are considered to provide a level of immersion and conditions for authentic communication that is difficult to replicate in traditional classroom environments (Dieterle & Clarke, 2008). Reinhardt and Sykes (2012) introduced the ways in which players participate in discourse with games which include *with the game*, as learners read or hear dialogue from game characters; *through and around the game*, as learners interact with other players while playing; and *about the game*, where learners engage in discussions about games after playing in non-game chat rooms and Internet discussion boards.

As well as the affordances for interaction and communication that VWs present for language learners and teachers, another catalyst for the current study is the apparent lack of empirical research on VWs in language learning contexts. Peterson (2010a, 2016a) notes that research on the use of 3D virtual worlds in the field of CALL is limited, and calls for more research on their applicability in language learning contexts. Furthermore, studies exploring the use of VWs in classroom-based instruction or framed from a TBLT perspective are even fewer. Whilst there are a number of papers that explore the conceptualisation of technology-mediated TBLT (Jauregi, Canto, de Graaff, Koenraad, & Moonen, 2011; Peterson, 2010a; Peterson, 2012; Milton, Jonsen, & Hirst, 2012; Sykes, 2014), an additional crucial area is still relatively unexplored: the effect of the environment on learners' oral interactions. Studies regarding VWs and language learning have focused on students' text-based interactions due to the affordance of the medium. That is, the dominant form of communication that occurs within VWs employs the written mode.

Additionally, in the CALL literature, there is often a focus on how L2 learners interact with native L2 speakers instead of their non-native speaking peers, again due to the networked nature of the medium (Rankin et al., 2008; Steinkhueler, 2006; and Thorne Black, & Sykes, 2009). Although such studies help to highlight the sociocultural and intercultural potential of autonomous language learning in VWs, the lack of studies positioned within monolingual L2 classrooms is rather unexplored. Finally, while previous studies have shown benefits of VW-mediated communication for both reading and writing skills (e.g., Chun,

2006; Oskoz & Elola, 2014), instructors still doubt the effectiveness of CALL activities for the promotion of L2 speaking and listening skills (Blake, 2016a). The current study explores the effectiveness of CMC, and in particular, the affordances of 3D virtual worlds for the development of speaking proficiency.

Another issue that remains unexplored is whether the additional cognitive load of interacting in a complex, virtual environment hinders novice-level EFL learners' noticing, remembering, and use of the L2. Kalyuga and Plass (2009) provide a thorough explanation of the brain's limited capacity for processing information from multimedia and games. They express concern that low-level learners may not have the cognitive resources to manage input from multiple lexical and graphical sources in games, quoting certain studies which utilised online games where learners were overwhelmed by the cognitive demands of learning tasks in addition to gaming (e.g., Lim, Nonis, & Hedberg, 2006). Cognitive overload from participation in an online game may be due to several factors, most prominently the steep learning curve required to understand how to play, and the specialised discourse used in such games. In addition, deHaan, Reed and Kuwada (2010) found that playing a video game hindered the uptake of vocabulary compared to non-playing "watchers" of the same game, a study emphasising that the extraneous cognitive load of a game negatively impacted learning. Related to this point, the Cognition Hypothesis (Robinson, 2007) attempts to categorise the complexity of language learning tasks and makes claims regarding how task complexity will affect learner output in terms of complexity, accuracy and fluency. Concerning the Cognition Hypothesis, it may be possible to "rank" the different tasks utilised in this study based on modality and task conditions. Learner oral performance can then be measured regarding two factors: modality (FTF and VW) and task complexity.

Few studies explore learner's spoken performance during tasks in VWs, and of those, task design is often left woefully unconsidered. Swier (2014) reviewed 14 studies to uncover how tasks had been designed for learners to undertake in VWs and revealed that the majority of studies only required learners to engage in open-ended computer-mediated discussions. In other words, studies to date tend to not fully utilise the affordances of the virtual world. Hampel writes:

Yet even today, the large majority of studies of computer-mediated communication

(CMC) – which are mostly concerned with the examination of written forms of communication and collaboration – deal with task design only tangentially and teachers frequently transfer tasks used in face-to-face settings to online environments without adapting them to the new setting. (2006, p. 103)

The aim of this study, then, is to investigate the usage potential of VW-specific tasks to improve novice EFL learners' oral proficiency as part of classroom-based instruction. This specific aim was created based on a perceived gap in the literature regarding this topic and the needs present in my teaching context. The learner participants in this study are low-proficiency Japanese university students who are all enrolled in an elective course designed to explore the potential of online language learning. While this experimental course may not be a typical language teaching contexts, one specific aim of the course is to improve students' speaking ability through various online and offline activities. The need for developing such communicative competence is also seen in a call from the Japanese Ministry of Education, Culture, Sports and Technology (MEXT) to nurture students that can "assertively make use of their English skills, think independently, and express themselves" (2014, p.3). Additionally, from a review of CALL, CMC, and technology-mediated TBLT studies, the current study aims to fill a gap in existing research by comparing learner oral production when undertaking tasks in a VW (computer-mediated communication) and face-to-face environments. This comparison is made following the literature on language learners' oral complexity, accuracy and fluency; an analysis method also known as the CAF model.

1.3 ORGANISATION OF THE THESIS

The outline of the thesis is as follows. The literature review (Chapter 2) starts by outlining the psycholinguistic underpinnings of TBLT, the chosen approach to classroom-based instruction utilised in this study and one of the predominant approaches to language teaching in general. Following this, the effects of task complexity and other task conditions on learner oral task performance are investigated. Such is conducted with reference to two prevalent theoretical constructs: the Cognition Hypothesis and Limited Attention Capacity Model which make claims regarding how task complexity affects learner performance.

The literature review also contains a detailed section on the genesis of digital game-based language learning (DGBLL), a sub-category of the computer-assisted language

learning field. This includes an overview of specific research paradigms related to language learning with games, based on the work of Sykes and Reinhardt (2012). The links between TBLT and DGBLL are also elucidated to position the current study as belonging to both a technology-mediated TBLT and game-based approach to language teaching.

Subsequently, I introduce and synthesise findings from relevant studies that utilise virtual worlds for language learning and teaching which are framed from a TBLT perspective. The final section of the literature review chapter is concerned with studies exploring learner output complexity, accuracy and fluency, particularly those that compare CMC and FTF based modes of communication. The chapter concludes with a summary of current findings and potential gaps in the literature and the research questions of this study.

Following the literature review, methodological considerations are presented in Chapter 3. This chapter introduces the rationale for choosing a mixed methods approach, an introduction to the participants and research context, as well as a detailed description of the VW employed and tasks designed for this study including the rationale for their inclusion. Data analysis methods are also introduced in the methodology chapter where I provide a rationale for adopting specific CAF measures and post-task questionnaire design.

Having introduced the research methods and data analyses, Chapter 4 presents the results of the study. Questionnaire data is analysed first in order to validate assumptions of task complexity manipulations. Upon completing this analysis, the effect of modality and task complexity on learners' output is introduced. A detailed discussion of findings is provided in Chapter 5 with reference to qualitative data collected in the form of learners' utterances during task performance and the open-ended questions of the post-task questionnaires. Finally, conclusions of this study are available in Chapter 6 where I summarise the findings in terms of their importance within the field of CALL, highlighting implications for other teachers that may be interested in employing a VW in their contexts. Chapter 6 also introduces the limitations of the study, which in turn provide suggestions for how the study may motivate future research.

2 LITERATURE REVIEW

2.1 INTRODUCTION

This chapter introduces the theoretical underpinnings of the current study in order to elucidate the specific approach adopted. This includes a review of studies which inform the design of this dissertation where a synthesis of findings is provided. The dissertation is framed from a cognitive-interactionist theory of SLA, with a particular focus on technology-mediated TBLT. As such, the first section of the chapter defines and outlines the importance of interaction in SLA, connecting Long's Interaction Hypothesis (1996) with tasks and TBLT. Multiple definitions of task are compared, including the specific definition adopted here, and then the specific pedagogical approach to TBLT is introduced with a focus on task design considerations.

Additionally, as task conditions and their effect on learners' output complexity, accuracy, and fluency are key constructs of this study, the two most influential cognitive-interactionist models of TBLT are introduced: Robinson's Cognition Hypothesis and Skehan's Limited Attentional Capacity Model. The two models are contrasted, and the differences that exist between them highlighted. Based on the claims made by the two models, the cognitive complexity of tasks designed in this study is calculated. Subsequently, the output complexity, accuracy and fluency (CAF) model is introduced including the myriad of available measures for evaluating learner's task performance. Findings from related studies are synthesised in order to highlight which measures are 1) prevailing in the literature and 2) appropriate for use in the current study.

Following an outline of the theoretical underpinnings of the study, the literature review then moves on to introduce the specific technology employed. A brief history of the use of virtual worlds for language acquisition is presented through the lens of CALL research, with ties to DGBLL also made. This helps to clarify that the current study is an intersection of technology-mediated TBLT, computer-mediated communication (CMC), and DGBLL.

Upon completing a detailed description of the technology and research areas to which this study connects, the literature review introduces studies that explore the effect of CMC on learner output, explicitly highlighting those studies which employ a VW as a domain for

interaction between learners. The literature review concludes with the formation of research questions based on a perceived gap in the literature and contributions that this study makes to the field are considered.

2.2 AN INTERACTIONIST MODEL OF SLA

Although Long writes that there is currently no dominant, unified theory of SLA (2014), within the field of instructed SLA (ISLA) there is a developing consensus around key parameters which help facilitate second language acquisition. One such parameter appears to be the role of *interaction*. Long's Interaction Hypothesis exists as part of a psycholinguistic approach to SLA, or, what is also known as a cognitive-interactionist theory of SLA. In the Interaction Hypothesis, then, Long (1983, 1996) proposes that interaction is a method of connecting input, internal learner capacities, selective attentional capacity, and output. During interaction, special status is assigned to the role of learner attentional shifts from meaning to form which occurs when there is a breakdown in communication, and *negotiation for meaning* is required (Long, 2014). The central thesis of the Interaction Hypothesis is therefore that interaction provides the means for the negotiation for meaning and as a result: language acquisition.

Breakdowns in communication may be signalled by the provision of implicit negative feedback in the form of clarification requests, confirmation checks, or recasts. In SLA, a recast is the immediate correct rewording of a learner's incorrect utterance. *Recasts* have been the focus of several studies, which have shown that they are the most common type of negative feedback inside classrooms (Lyster & Ranta 1997). Additionally, Long states that recasts have been found in every type of NS-NNS and NNS-NNS interaction studied to date (2014). Negative feedback and in particular recasts, therefore, indicate to learners that their utterances were problematic causing them to switch attention from meaning to form for brief episodes. From this, it has been hypothesised that learners may notice the formal qualities of their production (Swain, 1995), i.e. of the mismatch between their interlanguage and target-like forms of their interlocutors.

Accordingly, *noticing* one's errors is considered an essential part of second language acquisition and has a specific definition in SLA. Schmidt (1990) approached the concept of consciousness in language learning and distinguished three specific levels of awareness of

which *noticing* is the second. The first is "perception" which is the least conscious state and may not result in metalinguistic awareness (Long, 2012). "Noticing" follows, which requires a specific conscious awareness, making the noticed element(s) of input available for verbal report at a later stage. Finally, "understanding" is the ability to compare and analyse items that have been noticed on previous occasions, thus requiring the highest level of awareness.

Schmidt also notes that "more noticing leads to more learning" (1994, p.18). Upon receiving negative feedback, the learner may modify their output in order to repair and resume interaction. Thus, learners are likely to switch their attention from meaning to form when there are problems with communication. This purposeful switch provides learners with the time to solve communication problems and notice any necessary new information (White, 1987).

There have been numerous empirical studies on the Interaction Hypothesis over the last three decades with findings showing that interaction that occurs as part of undertaking tasks has a positive effect on a range of morphosyntactic features. Mackey, Abbuhl and Gass (2012) provide a detailed overview of such findings, which include the acquisition of articles (Sheen, 2006), question formation (Philp, 2003), past tense formation (Ellis, 2007; McDonough, 2007), and plurals (Mackey, 2006). Gass and Mackey (2007) also state that "it is now commonly accepted within the SLA literature that there is a robust connection between interaction and learning" (p. 176).

The creation of the Interaction Hypothesis coincides with the decline of grammar-translation and behaviourist approaches to language learning (for instance, audiolingualism) and a shift towards more communication-driven approaches such as communicative language teaching (CLT). For instructed SLA contexts, then, promoting learners to interact in the L2 is seen to have a positive influence on their language acquisition, and is backed up by both theory and research. However, how one promotes interaction that is beneficial to L2 development is related to appropriate pedagogical implementation. It is to this topic which the literature review now turns, as specific tasks as part of a TBLT approach to instructed SLA are considered particularly useful in promoting interaction for successful L2 acquisition.

2.3 TASK-BASED LANGUAGE TEACHING

The status of *task* as an important facilitator of L2 development is undeniable,

particularly in instructed SLA settings. Tasks, regarded as interactive communicative activities, are considered beneficial in promoting interaction between learners. Additionally, benefits include promoting learners to see language as functional (Bygate, Norris, & Van den Braden, 2015) and improving learners' interlanguage through hypothesis testing and negative feedback during task performance (Swain & Lapkin, 1998; Long, 1996). Subsequently, TBLT, which features tasks as the core unit of analysis in syllabus design, may be considered the most researched pedagogical approach to instructed SLA (Long, 2014). Indeed, its popularity as an approach has captured the attention of teachers, researchers and materials developers alike (Sasayama, 2015). Proponents of the approach generally agree that tasks are "pedagogically useful, practically relevant, and psycholinguistically valid" (Feryok, 2017 p. 717). However, regardless of the amount of research and data collected on TBLT, its application in classroom contexts is not without difficulties. TBLT is an often-misunderstood concept (Carless, 2004; Ellis, 2009), perhaps because of the multiple definitions of task that exist, and the differing opinions on how to appropriately "do" TBLT appearing in the literature (Willis & Willis, 2007; Long, 2016). The following section focuses on the former issue: the definition of tasks, before outlining the particular pedagogic approach to TBLT adhered to in this dissertation.

2.3.1 Task: a definition

In order to establish task design considerations for this study, a definition of *task* is needed. Long (2014) provides a succinct overview of the range of definitions from the "non-technical everyday real-world use of the term" (p. 108) to those which he calls "abstract and opaque" (p. 109). For instance, Candlin describes a task as:

> One of a set of differentiated sequence of all problem posing activities involving learners and teachers in some joint selection from a range of varied cognitive and communicative procedures applied to existing and new knowledge in the collective exploration and pursuance of four for seen a or imagined goals within a social milieu. (Candlin 1987, p. 10)

Long's definition is that tasks are "the real world activities people think of when planning, conducting, or recalling their day" (p. 6). Following, he describes such tasks as "target tasks" (p. 109), which are identifiable as the tasks that learners will be required to

complete in the L2 based on a needs analysis. Following this description, Long makes a distinction between target tasks and "pedagogic tasks." In other words, the tasks created by educators as stepping-stones to achieving the identified target tasks. Keeping a focus on target tasks, he provides two additional, similar definitions. The first is by Crookes (1986, p. 1) who defined a task as a piece of work or activity "usually with a specified objective, undertaken as part of an educational course or work." Second is Skehan's (1998) definition: "Meaning is primary; there is a goal which needs to be worked on; the activity is outcome-evaluated; there is a real-world relationship."

Alternative definitions for a task have been conceived by scholars such as Ellis (2003), Nunan (2004) and Willis and Willis (2007). Most definitions of task have similar properties such as:

1) a primary focus on the provision of meaning between learners as they engage in language use,
2) non-linguistic goals, which require the use of authentic language in order to be completed, and
3) that tasks resemble activities that learners do with language outside of the classroom, in other words, a real-world use case for the language spoken during task performance.

Sociocultural theory adds a further dimension to the description of a task, which includes participants' individual goals. That is to say, task goals exist in two forms; 1) as how the instructor intended (task-as-workplan), and 2) how the participants interpret task goals based on their socio-history (task-as-process) (see Coughlan & Duff, 1994; Breen 1989). Appel and Lantolf (1994) add, "performance depends crucially on the interaction of individual and task" (p. 437). Thus, the distinction of a task for SCT researchers is an activity that may have an initial instructor-defined goal, but may be perceived differently, and therefore carried out differently even if participants undertake the same task in the same context and with the same resources (Yuksel, 2003). As an attempt to quell the seeming uncertainty that sociocultural theory brings to the notion of task-goal interpretation, psycholinguists reconcile the view that the linguistic output of learners during task performance may be predicted. Ellis (2003) defines "focused tasks" which are those tasks

explicitly designed to target particular grammatical features, thus providing instructors with the ability to predict the linguistic output of learners.

For language teachers, it is considered natural to define task from the perspective of what happens in the classroom. Therefore, in the current study Samuda and Bygate's (2009) definition of task was chosen because it is succinct, concrete, and makes direct reference to language learning. "A [task is a] holistic activity which engages language use in order to achieve some non-linguistic outcome while meeting a linguistic challenge, with the overall aim of promoting language-learning, through process or product or both" (p. 69).

However, how can tasks be designed to promote interaction between learners? Pica et al. (1993) devised a task typology for choosing communication tasks based on the Interaction Hypothesis. The logic follows that if interaction is beneficial in providing opportunities for learners to receive comprehensive input, feedback on their production, and interlanguage modification, then choosing tasks that best promote these features is paramount.

2.3.2 Task typology

Pica et al.'s typology can be seen in Table 1. The table shows the relationship and communication requirements between two interactants: X and Y. *Information holder* refers to which of the two interactants holds the information required to complete the task. *Requester* and *Supplier* refer to which of the two interactants are required to request and supply such information. The *relationship* between the interactants is considered either one-way or two-way. Interaction may not be required for the successful completion of specific tasks, and so is considered a unique feature of the typology. Subsequently, *goal orientation* may be more or less convergent (±), and finally, tasks may have one or more *outcomes*.

Table 1: Task typology for communication task (Pica et al. 1993, p.18). INF = information.

Task type	Inf Holder	Inf Requester	Inf Supplier	Inf Requester-Supplier relationship	Interaction requirement	Goal orientation	Outcome options
Jigsaw	X & Y	X & Y	X & Y	2 way (X to Y & Y to X)	+ required	+ convergent	1
Information gap	X or Y	X or Y	X or Y	1 way > 2 way (X to Y / Y to X)	+ required	+ convergent	1
Problem-solving	X = Y	X = Y	X = Y	2 way > 1 way (X to Y & Y to X)	- required	+ convergent	1
Decision-making	X = Y	X = Y	X = Y	2 way > 1 way (X to Y & Y to X)	- required	+ convergent	1 +
Opinion exchange	X = Y	X = Y	X = Y	2 way > 1 way (X to Y & Y to X)	- required	- convergent	1 ±

For *jigsaw* tasks, both participants hold, request and supply information to their interlocutor in order to achieve a single, convergent task goal. This type of task is considered the most likely to promote interaction between participants as they pool their information together; thus, the most opportunities for successful SLA development. *Information gap* tasks are similar to jigsaw tasks, but the information exchange is one-way. This means that one of the interactants holds all of the required information for task completion and supplies it to their interlocutor. However, if the task is then repeated, reversing roles, two-way communication may be established.

For the following three task types, *problem-solving*, *decision-making* and *opinion exchange*, both of the participants have access to the same information at the start of the task, thus the notion of X = Y in the table. Two-way communication is possible, but not always necessary. This refers to how if both participants have access to the same, shared information

at the start of the task, they may work individually towards achieving the task goal such as solving the problem, making a decision, or formulating an argument for the opinion exchange. Additionally, as interaction is not required (-), one of the participants may dominate the conversation leading to one-way communication.

Problem-solving tasks may promote interaction as participants work towards mutual understanding of the problem at hand, and the possible solution from the provided information, but not necessarily. The sharing of information between participants means that such interaction is not guaranteed. Goal-orientation for decision-making tasks is not singular so participants may seek one of many possible decisions, exchanging information (or not) as part of the process in reaching their decision. Finally, *opinion exchanges*, as the name suggests, requires participants to exchange their ideas, not reach a consensus (- convergent). Thus, this task type may end with participants verbalising their take and inevitably holding their original stance regarding the issue of the task.

In summary, then, by creating a task typology, Pica et al. (1993) were able to hypothesise that interaction between learners can be promoted to varying degrees based on seven task components. They also proposed that jigsaw tasks provide the most opportunity for interaction, with information-gap activities a close second depending on whether the task is repeated with roles reversed.

Having introduced the importance of tasks from a cognitive-interactionist perspective, and how various task types may afford different levels of interaction between learners, the next section turns to the broader pedagogical considerations of TBLT and how it may be implemented in instructed SLA contexts.

2.3.3 TBLT framework

The approach to TBLT employed in this study is informed by Klapper (2003) and Ellis (2003) in what is known as a 'weak version' of TBLT. In this version, the communicative interaction characteristics of tasks are considered essential to providing comprehensible input to learners, and thus triggering language acquisition. However, compared to the strong version of TBLT (what Long refers to as TBLT with capital letters) the weak version of TBLT employed here recognizes that explicit instruction of linguistic forms can help facilitate acquisition after fluency-focused activities in what is known as

"focus on forms" (Doughty & Williams, 1998; Doughty, 2001). The general model which informs the pedagogical considerations of this study is provided in Table 2. This framework, which is based on Willis' (1996) well-established and teacher-friendly model of TBLT, is comprised of three distinct sections: the pre-task stage, the task cycle, and the post-task stage.

Table 2: An example TLBT lesson framework

Stage	**Sub-stage**	**Activities**
Pre-task		Learners are prepared for the task through relevant discussion, brainstorming, or other appropriate activities.
Task		Learners carry out the non-linguistic task.
Post-task	Report	A report (spoken or written) is produced reflecting on the completed task.
	Focus on forms	Teacher-led form instruction.

The pre-task stage, also known as the priming stage, may be considered a task itself as students listen to the teacher introduce the topic of the lesson. The teacher may highlight useful phrases and words or introduce a recording of a native speaker carrying out a task similar to the one designed for the upcoming task-phase. Typically, this stage of the lesson is teacher-led; however, students may be asked to brainstorm words or phrases that they think might be useful during the upcoming task cycle.

The task cycle is the stage of a lesson where students work together in pairs or small groups to complete a meaning-based task that has a direct relation to the topic introduced in the pre-task stage. Students are focused on meaning during the task as they communicate with their peers in an information gap or opinion-based activities.

In the report phase, learners must report on what they have discovered or how they completed the task. This requires the use of accurate language; thus, group members must work together to focus on not only what they plan to say, but also how they will say it. Thereby, a student-led focus on not only language fluency, but also accuracy is realised in this part of the lesson. The final stage of Willis' lesson sequence is known as "focus on forms," and consists of language focus activities that encourage students to analyse the language that they have been using during the task phrase. This specific stage can be considered an extended, formal, teacher-led version of the brief switching of focus that occurs

during task time interaction. In other words, as learners switch their attention from meaning to form during task time, they are engaging in explicit form-focused learning. However, some researchers have raised concerns as to whether this short attentional shift is enough for learners to acquire grammatical forms, particularly in the Japanese contexts (Burrows, 2008; Sato, 2010). As a result, a stage for focusing specifically on form after completing a task is considered essential (for a summary, see Hawkes, 2011). The stage is often teacher-led, with premade materials based on the target grammar the task is predicted to promote or, reactively, based on errors in learner output that the instructor observed during task performance. The above considerations were closely followed when designing the pedagogical approach to this study. Each lesson features a pre-task, task, and post-task phase as a psycholinguistically sound approach to aid learners in their L2 acquisition.

2.3.4 Measuring task complexity

There is a large volume of research exploring the effect of task conditions on task performance, of which two influential models help frame the discussion: Robinson's Cognition Hypothesis (2001) and Skehan's Limited Attentional Capacity Model (1998). This section introduces the concept of cognitive task complexity, a detailed description of task conditions, and a comparison of the two models. The effects of manipulating task conditions on learner output are introduced later in the chapter after the concept of linguistic complexity, accuracy, and fluency as well as the technology employed in the study. This is to allow for a specific review of studies which explore the effect of task condition manipulations on learner output when conducting computer-mediated communication tasks.

2.3.4.1 Skehan's Limited Attention Capacity Model

There are two models which help us predict how a specific task will affect learner output. The first model was developed by Skehan (1998, 2001, 2003) and is known as the Limited Attention Capacity Model (LACM). It is based around a concept from cognitive psychology which states that humans have limited working memory and limited attentional capacity, governed by a single, central control mechanism (Skehan, 2009). Skehan created the LACM based on observations of higher accuracy and fluency in learner output when performing specific tasks (Foster & Skehan, 1996; Skehan & Foster, 1997). The model is also underpinned by VanPatten's (1996) Input Processing Theory which also found that

learners (particularly low-level learners) struggled to attend to both meaning and form during a listening task.

Content and form are hypothesised to compete for attentional resources, and, as humans are mostly concerned with meaning over form, as tasks increase in complexity, learners may tend to focus more on the meaning of messages over the words needed to express themselves. The result is that due to limited attentional resources, only certain aspects of oral proficiency (complexity, accuracy, or fluency) are focused upon at the expense of others. This is known as *the trade-off effect*.

Skehan (1998) has suggested that the cognitive complexity of tasks may be considered to consist of three different components: code complexity, cognitive complexity and communicative stress. *Code complexity* arises from the inherent difficulty of linguistic forms required to complete the task. This also relates to the developmental stage of the individual learner, as one learner may perceive the linguistic difficult of a task to be higher than a more proficient interlocutor. *Cognitive complexity* is concerned with the content of the task and can be further divided into *processing* and *familiarity*. Processing refers to the level of online processing a learner must do as part of the task (use of short-term memory, manipulation of data, inference, and calculation). A learner's familiarity with the schema of a task may affect cognitive complexity where higher familiarity will reduce cognitive complexity. Knowledge of how to complete the task at the outset thus reducing the amount of processing required. The third construct, *communicative stress* relates to aspects of a task that are not directly related to linguistic form or meaning. Examples include time pressure, modality, number of participants, stakes (i.e. pragmatics, face, the importance of the task), and level of control (i.e. whether a participant can negotiate task goals).

Skehan (2001) offers the following generalisations found in empirical studies of task complexity and its effect on output complexity, accuracy and fluency in learner proficiency. These are:

- tasks based on familiar information promote **accuracy and fluency**;
- clearly structured tasks promote **accuracy and fluency**;
- interactive tasks promote **accuracy and complexity**;
- information manipulation may lead to **higher complexity**;

- Post-task conditions such as public performance or transcribing one's performance may promote **higher accuracy.**

As can be seen in the above list, task conditions are generally considered to promote increased attention towards two aspects of oral proficiency, where a single aspect is traded-off. For example, with clearly structured tasks, learners are predicted to give attentional priority to accuracy and fluency at the expense of complex output.

In sum, the LACM states that an increase in task complexity can help produce an increase in performance along one linguistic dimension (accuracy, fluency, or complexity), but at the expense of the other two (e.g., an increase in linguistic complexity will result in a decrease in accuracy and fluency).

2.3.4.2 Robinson's Cognition Hypothesis

There is a complementary model to Skehan's LACM, which is Robinson's (1995, 2001a, 2011b, 2007) Cognition Hypothesis (CH). This model provides a framework for categorising task conditions into three broad categories: *task complexity* (based on cognitive factors), *task condition* (based on interactive factors), and *task difficulty* (based on learner factors). This framework was known as the Triadic Componential Framework (Robinson, 2007), but has since been renamed to the SSARC Model (2010). SS = simple/stabilizing interlanguage; A = automatizing access to interlanguage; and RC = restructuring and complexifying.

Table 3: Robinson's Triadic Componential Framework for task conditions (2007).

Task complexity		Task conditions		Task difficulty	
Resource-directing	**Resource-dispersing**	**Participation**	**Participant**	**Affective variables**	**Ability variables**
+/- Here-and-Now	+/- Planning	+/- Open solution	+/- Same proficiency	H/L Openness	H/L Working memory
+/- Few elements	+/- Prior knowledge	+/- One-way flow	+/- Same gender	H/L Control of emotion	H/L Reasoning
+/- Spatial reasoning	+/- single task	+/- Convergent solution	+/- Familiarity	H/L Task motivation	H/L Task switching
+/- Causal reasoning	+/- Task structure	+/- Few participants	+/- Shared content knowledge	H/L Processing anxiety	H/L Aptitude
+/- Intentional reasoning	+/- Few steps	+/- Few contributions needed	+/- Equal status and role	H/L Willingness to communicate	H/L Field Independence
+/- Perspective-taking	+/- Independency of steps	+/- Negotiation not needed	+/- Shared cultural knowledge	H/L Self-efficacy	H/L Mind-reading

What separates Robinson's model from Skehan's is the *multiple resource* view of attention compared to Skehan's *limited capacity model* (see Robison, 2003). Thus, whereas Skehan's LACM has only one category for task complexity, the CH model distinguishes between *resource-directing* and *resource-dispersing* conditions (see Table 3, above). The *complexity* of a task is determined by the cognitive demands placed on learners and exist inherently as part of tasks as designed by researchers or instructors. There are several criteria given to task complexity from whether the task is performed in the here-and-now, how many elements learners are required to interact with, several different reasoning elements such as spatial, causal and intentional reasoning and the amount of planning time learners receive.

Robinson predicts that increasing task complexity along the *resource-directing* dimensions of tasks may have a positive effect on complexity and accuracy at the expense of fluency. However, regarding *resource-dispersing* dimensions such as depleting planning time or requiring learners to focus on multiple simultaneous tasks at once, he predicts that accuracy and complexity of production can be expected to decrease as task complexity increases. Compare this to Skehan's model which sees task complexity along any dimension as negatively affecting performance in one of the three attributes of proficiency: fluency, complexity or accuracy (Robinson & Gilabert, 2007).

In more detail, the claims of the CH are that increasing the cognitive demands of tasks

along the resource-directing dimension will:

- **push learners to greater accuracy and complexity of L2,**
- promote interaction and heightened attention to input, thus increasing learning from the input,
- longer-term retention of input,
- sequencing tasks from simple to complex will lead to automaticity and efficient scheduling of the components of complex L2 task performance (based on Robinson & Gilabert, 2007).

The bolded text above is of particular importance to this study, and Robinson (2001b) adds to this claim that there is also a difference in performance for monologic and dialogic task conditions. This is in opposition to the claim of Skehan that "*interactive tasks* promote accuracy and complexity." According to Robinson, monologic conditions are said to promote syntactically complex and accurate, but less fluent output. However, under dialogic conditions, the assumption is that syntactic complexity is reduced due to the contributions that interlocutors make. Additionally, tasks that make use of complex spatial reasoning are expected to promote the use of more advanced lexical patterns to describe events related to motion (Cadierno, 2004). This point has relevance to the current study because of the requirement of learners to perform tasks with reference to materials in two different spatial contexts – 2D (face to face) and 3D (VW). We may therefore expect the VW tasks to promote more complex language use due to the increased spatial complexity of the domain.

Regarding the task complexity condition "*number of elements*", Skehan (2016) conducted a meta-analysis of studies that explore the effects of the number of task elements on learner output complexity, accuracy, and fluency. According to this meta-analysis, there appears to be little connection between the number of elements (i.e. task complexity) and the complexity of learner output. Only one of the ten studies outlined suggested that there was a positive link between task complexity and learner spoken complexity (Sasayama & Izumi, 2012). Additionally, due to the large volume of variables in the CH model, it is considered difficult to operationalise and isolate them (D. Ellis, 2011). An example given in Long (2014) is that it is even difficult to conduct an experiment based on the seemingly transparent variable "number of elements." This is due to a possible mismatch in researcher and learner

perceptions of how many elements were included in a task. If a researcher claims to have included ten elements in a test, but only five of those were salient to the test subject, there is no reliability in terms of this variable.

The +/- here-and-now demand also appears to have a significant influence on learner output. Studies include Robinson (1995) and Rahimpour (1999) who elicited participants to describe a cartoon strip in a classroom setting and Iwashita et al. (2001) whom based picture prompts on the Test of Spoken English where participants carried out the task in a laboratory setting. Results tended to show that tasks conducted in the +here-and-now elicited greater fluency but lower accuracy on narrative performance. However, the tasks used in these studies were manipulated along the +/- here-and-now dimension, and also required participants to use different tenses (present tense for the + here-and-now version and past tense for the – here-and-now version) meaning that a metalinguistic demand was also placed upon them. Concerns have been raised about the comparability of such studies, as the effect of providing and removing visual stimuli and production in the present and past tense are invariably linked. In relation to the current study, the results of the meta-analysis indicate that care must be taken to ensure that resource directing demands are carefully controlled. This is particularly important along the +/- here-and-now condition.

Sasayama (2015, p.36) provides an overview of studies which have operationalised task complexity to be either more or less complex. Manipulation of task complexity is thought to affect learners' task performance or task success due to differing cognitive demands placed on them. However, Sasayama posits that there are at least three issues with this assumption (p.37). First, task conditions have differing levels of inherent complexity. That is, manipulating one task condition may only produce slight changes in overall task complexity whereas manipulation of another condition could produce more significant changes to task complexity. Secondly, the particular cognitive demands posed by each condition are not comparable, i.e. (+/-) planning time (a pre-task condition) may pose different cognitive demands to (+/-) number of objects (a during-task condition). Moreover, the manipulation of task conditions alone may not be the source of differing task performance. This final point positing that learner factors and pedagogic intervention may have more of an effect than any single task condition (see also Skehan, 2016).

Compared to the large volume of research that exists on task complexity manipulations and their effect on learners performance in face-to-face dialogic interactions there has been less research on how modality or differing levels of task complexity as part of CMC may affect learners' performance. This is especially true of VW-based SCMC, which the current study explores.

Task Difficulty is a separate construct, based on individual learner differences, such as their current L2 proficiency, vocabulary range, reading skill, background knowledge related to the task and so on. That is to say, the difficulty of a task is a combination of the inherent, implicit complexity, and the ability of individual learners. This relates somewhat to the LACM construct of both *code complexity* and *familiarity* with a task. From the perspective of the current study, technological proficiency may be considered one substantial determiner of learners' perception of task difficulty.

2.3.4.3 Model comparisons

Ellis (2000) provides a summary of studies conducted on both the LACM and CH models and concludes that it is unrealistic to favour one model over the other due to a lack of cohesion regarding the tasks used. Some studies have tried to forcibly align their results with one of the two models (such as Gilabert, 2007) yet there are several problems with this. For instance, some studies employed monologic formats (Yuan & Ellis, 2003), others dialogic (Gilabert, Barón & Llanes, 2009), and further still, some have featured both (Michel, Kuiken & Vedder, 2007); studies varied in the level of detail regarding task description; complexity variables have been used in different combinations and manipulated to different extents within the same variable condition; and studies used spoken, written, or both modes of communication (see Jackson & Suethanapornkul, 2013 for a review).

Furthermore, the methods of data analysis employed in studies exploring task complexity and learner output vary considerably. Although there is a large number of empirical studies that explore the relationship between task complexity and CAF (for example Cadierno & Robinson 2009; Robinson 2011; Robinson, Cadierno & Shirai, 2009), there is not a single method for analysing CAF data. Jackson and Suethanapornkul (2013) identify 84 different approaches to data analysis in CH studies alone. Additionally, the many options of data analysis for spoken, and to some extent written data, can be seen in Norris

and Ortega (2000), and Ellis and Barkhuizen (2005).

A simple difference between the two models is in the conceptualisation of task complexity. As mentioned above, Skehan does not distinguish between resource directing and dispersing dimensions as Robinson does, but claims that one area of performance over the other two will be focused on depending on task conditions. This leads to an ambiguous conclusion where any of the three categories of performance (complexity, accuracy, or fluency) could be positively or negatively affected by task complexity. Additionally, Skehan (2009, 2016) argues that a crucial difference between the two models is that in addition to task complexity, pedagogical interventions may be more influential on learner performance. Examples given are the amount of planning time learners receive, and whether this planning time is teacher or student led. Thus, factors outside of a task's design (and therefore task complexity) may have more of an effect on learner output performance.

Finally, Sasayama (2015) writes that a discrepancy between the two models is whether linguistic complexity and accuracy can be attended to simultaneously or not. However, claims regarding both models are that certain manipulations of task conditions may promote an increase in accuracy and complexity. The difference is in how both models predict this is produced. For the LACM, interactive tasks are considered beneficial in promoting accuracy and complexity, whereas for the CH increasing the cognitive demands of tasks along the resource-directing dimension is thought to push learners to greater accuracy and complexity.

2.3.4.4 Summary of task cognitive complexity effects on learner output

There are two prevalent models for predicting how learner output is affected by task complexity. These are the LACM (Skehan) and Cognition Hypothesis (Robinson). Although several meta-analyses have been conducted on the effect of task type on learner output, neither model can currently be considered the prevalent model for predicting language output. Robinson's model is unique in that it separates task *complexity* into several dimensions and posits that resource-directing demands such as the (+/-) here and now resource is particularly influential in determining learner output. Skehan (2016) however writes that the impact of pre-task planning, task repetition, and post-task activities on learner

output may be more significant than the impact of tasks and features such as the number of elements (therefore pushing for the validity of his own model over the CH).

Thus, task complexity and its effect on learner output is a complicated and largely contested topic. In terms of the current study, it is possible to make predictions regarding the complexity of different tasks based on task conditions such as +/-*here and now*, +/-*number of elements* and so on. Additionally, based on Skehan's work (2016), due to the potentially considerable influence of pedagogical interventions pre and post-task, care should be taken to keep such activities as uniform as possible for all tasks in the experiment.

The above section outlined the two predominant models for considering the complexity of pedagogic tasks, and how different tasks may affect learner output in terms of complexity, accuracy and fluency. The following section introduces the various methods of measuring learner output performance that appear in the literature.

2.3.5 Output complexity, accuracy, and fluency: A definition

Many researchers now hold that L2 proficiency is multifaceted and that the CAF model can help to capture the main elements of this proficiency through the collection and analysis of both quantitative and qualitative data (e.g. Ellis 2003, 2008; Housen & Kuiken, 2009). As such, the CAF model has been appearing in the literature more and more in recent years alongside traditional proficiency models such as the four-skills model (listening, speaking, reading and writing) and sociolinguistic model (e.g. Bachman 1990; Bialystok 1994; Canale & Swain 1980).

The notion of accuracy and fluency first appeared as a dichotomy in Brumfit (1984), who used the terms as descriptors for activities that were predisposed to promote either accurate or fluent language use among learners. The origin of the CAF model, however, can be traced back to the work of Skehan (1998) who proposed an L2 model comprised of the three different proficiencies. Since the founding of the CAF model, there have been a plethora of measures introduced for assessing learner performance. Due to the large volume of measures available, careful consideration is needed in choosing appropriate measures for the current study. The following section outlines several measures that are used to measure each component of the CAF model.

2.3.5.1 Complexity measures

Jackson and Suethanapornkul's (2013) meta-analysis on task complexity reveals a number of measures that have been used for assessing learners' output complexity. Measures in their "general measurements" classification (p. 347) are presented in list form below. An explanation of key constructs follows.

- Clauses per
    - AS-unit
    - C-unit
    - T-unit
- Dependent clauses per clause
- Multipropositional utterances
- Ratio of dependent clauses to total clauses
- S-nodes-per-clause
- S-nodes-per-T-unit
- Subordinate/total clauses
- Total clauses
- Words per turn

It is worthwhile noting the different units used to measure performance. *AS-Unit*, *C-Unit*, *and T-Unit* are all represented in different studies, often based on the mode of communication that the learners are engaged in. For example, monologic, written performances promote the use of longer utterances than dialogic, spoken performances. In this way, there are specific units designed for each situation.

The *t-unit* was initially devised by Hunt (1965) and is defined as an independent clause together with all its dependent clauses. The "t" is an abbreviation of "terminable." Often, the t-unit is considered a sentence. The t-unit is generally used to assess learners' written prose; however, it has also been used with spoken performances also. Its equivalent is the *communication* unit or *c-unit,* which was developed specifically for oral communication (see Foster, Tonkyn, & Wigglesworth, 2000). The c-unit is defined as a single complete sentence, phrase, or word that has a clear, pragmatic relevance in the context it is used (Johnson & Johnson, 1999). Finally, the AS-unit (Analysis of Speech Unit) is

defined as a "single speaker's utterance consisting of an *independent clause* or *sub-clausal* unit, together with any *subordinate clause(s)* associated with either" (Foster et al., 2000 p. 365). This unit is primarily designed for the analysis of spoken language as evident by the addition of the "sub-clausal unit" in the definition.

As the current study is concerned with learners' spoken performance, it may be beneficial to use the AS-Unit over the t-unit. Indeed, the unit appears to be utilised in studies where learners are engaged in oral production such as in Sample and Michel (2014), who investigated the relationship between task repetition and young learners' oral complexity, accuracy, and fluency.

Jackson and Suethanapornkul (2013) differentiate *complexity* from *lexis*, reclassifying the model acronym as "CA*L*F" instead. For lexis then, they note the following measures. Of these, I will focus on the measures that appear most abundantly in the literature:

- % Lexical words
- Guiraud's Index (a measure of lexical richness which is defined as the total number of words / square root of the total number of tokens)
- Guiraud 2000 (similar to Guiraud's Index for spoken language considering only words that are within the 2000 word frequency list)
- Ratio of lexical to function words
- Ratio of lexical to total words
- Ratio of word types to square root of 2 * number of tokens
- Token type ratio
- Type token ratio
- Word types squared/total # of words
- Word types/square of 2*total # of words

The type-token ratio (TTR) is possibly the most common measure of lexical complexity (Koizumi, 2012) and has been utilised in studies of both spoken and written

performances (Revesz, Ekiert, & Torgersen 2014; Kuiken & Vedder, 2006). The type-token ratio is calculated by dividing the *types* (or, the total number of different words) by the *tokens* (the total number of words) in a text or utterance. Thus, a high TTR value equates to a high degree of lexical variation. The range of TTR goes from zero to one. However, the accuracy of TTR has been shown to vary substantially based on the length of texts. As a replacement, "D," has become a viable alternative to TTR (Malvern & Richards, 1997). The D-value is calculated by the probabilistic mathematical model using a random sampling of tokens in calculating the type-token ratio (Kormos & Trebits, 2012). It is thus statistically controlled for different text lengths (Revesz et al., 2014). D, then, does not depend on the length of samples in order to be used, making it a suitable measure for shorter samples. Additionally, there is a third commonly used measure for lexical density that, again, appears as a successor to the TTR and is of considerable importance to the current study. It is known as the 'measure of textual lexical diversity' (MTLD). Koizumi (2012) researched which of four lexical density measures was least affected by text length with a focus on short L2 texts (50-200 tokens). The measures were: TTR, Guiraud, D, and MTLD. Results of the study suggested that MTLD was least affected by text length, but that it should be used with texts of at least 100 tokens. This has relevance to the current study in that I am unable to predict the typical length of learner output as they complete tasks. Upon collecting data, one of the above four measures will be employed to calculate learners' lexical complexity.

Finally, although I have provided details of an extended CAF model with lexical complexity separated from general/syntactic complexity (making the abbreviation: CA*L*F), I will continue to use the abbreviation and categorisation of "CAF" in this study, including lexical complexity as a part of complexity.

2.3.5.2 Accuracy Measures

Accuracy may be considered similar to the word "correctness" in that it refers to how much an utterance deviates from the norm (Housen, Kuiken, & Vedder, 2012). However, it is debatable how the term "norm" should be defined. Possible definitions include a comparison of learner utterances compared to native speakers, to other non-native speakers, or the learner's output at varying stages of development (see Ågren, Granfeldt, & Schlyter, 2012). Of the CAF triad, accuracy is considered the most transparent construct, and regardless of how the term "norm" is defined, diversions from the norm are characterised as

errors.

The most common measure for accuracy was defined by Skehan and Foster (1996) as the overall percentage of error-free clauses of learner utterances during oral tasks. In a later paper, they considered this generalised view of accuracy as "sensitive to detecting significant differences between experimental conditions" (Skehan & Foster, 1999 p. 229), i.e. in comparing the effect of varying task conditions on learner accuracy. Ellis and Yuan (2003) adopted the same definition of accuracy for their study but provided further details in the form of multiple error types: *syntactic*, *morphological*, and *lexical* errors. Note that pronunciation was not included.

Ellis (2009) conducted a meta-analysis of studies exploring the effects of planning time on learners' oral complexity, accuracy and fluency. Focusing on accuracy measures, the following additional examples are provided. However, they appear less frequently in the literature:

- errors per t-unit (Bygate, 2001)
- percentage of correct verbs (Wendel, 1997)
- number of errors per 100 words (Guara-Tavares, 2008)

In summary, then, there appears to be fewer measures for assessing learner accuracy than those for complexity. One definition, in particular, is used most often: error-free clauses, where errors are subdivided into morphological, syntactic and lexical errors.

2.3.5.3 Fluency measures

The previous two sections showed that there are a large number of measures for complexity and a more limited number for accuracy. Fluency, like complexity, is also considered a complicated concept that can be measured in various ways. There are however two dominant types of output fluency: temporal and vocal fluency.

- Temporal fluency, which includes
    - Rate of speaking (Ejezenberg, 2000; Kormos & Denes, 2004)
    - Length of fluent runs before a pause (Kormos & Denes, 2004)
    - Frequency, length and placement of pauses based on Skehan's (2009) *breakdown* and *speed* fluencies

- Vocal fluency, which is indicated by
    - Number of false starts
    - Filled pauses (Lennon, 1990; C. Blake, 2009)
    - Reformulations
    - Functionless repetitions, which is equivalent to Skehan's (2009) *repair* fluency.

C. Blake (2009) cites Sajavaara and Lehtonen (1978) as researchers who have argued that fluency is too complex to be assessed with only a few temporal variables. However, rate of speech appears frequently in the literature as the sole measurement of learner fluency (e.g. Geng & Ferguson, 2013; Yuan & Ellis, 2003; Wendel, 1997). One possible reason for this is due to the ease of assessment via quantitative analysis. Researchers also state this as one of the primary reasons for selecting it in their study (e.g. Geng & Ferguson 2013, p. 984) "for ease of reporting we focus on a single measure: pruned speech rate; that is, words per minute excluding false starts and field pauses." However, though rate of speech is one of the more straightforward fluency measures to implement does not mean that it is without value. It is an accurate measure of fluency, and appears in several studies (Kormos & Denes, 2004; Kormos & Trebits, 2012). Grant and Ginther (2000) also note that within the written domain, filling time with comprehensible sentences correlated highly with a participants' perceived language ability.

In a study by Sample and Michel (2014) which explored the effect of task-repetition in young learners' oral production, fluency was also analysed based on the time it took dyads to complete two instances of the same task. The notion being that fluency is related to the speed at which learners complete a task. For the current study, however, although tasks in a task pair could be considered "repeated" instances, vocabulary required for task completion differs per mode. That is, although the tasks in a task pair are matched to be the same task type and have the same task goals, specific affordances (or limitations) of each mode resulted in tasks being designed differently, taking into account these affordances. The measure was thus not considered appropriate. As a concrete example, due to the nature of the VW tasks, player actions are predicted to take longer within this modality and, therefore, as a result, task completion time is predicted to be longer than the FTF task equivalents.

2.3.5.4 Summary of the CAF model

In summary, there is a large variety of measures for assessing the complexity, accuracy and fluency of learner output. For the current study, measures should be chosen based on a critical review of those used in previous, related studies. The selection was made through assessment of two factors 1) how other studies employed specific measures, and 2) how similar those studies are to the current study. By favouring measures that already appear in the literature, it allows for the comparison of any findings made here with those of previous studies and avoids diluting the range of analysis techniques any further. The measures chosen are elaborated upon in the methodology chapter.

2.3.6 Open and closed task definition

Due to the results of a pilot study (see Section 3.4.1), the current study only focuses on the use of closed-goal tasks. It is therefore essential to outline what is meant by this definition by contrasting it to the other task-type: those with open-goals (henceforth referred to as closed tasks and open tasks). Long (2014) defines closed tasks as a task which "require students to find the correct solution, or one of a small, finite set of correct solutions to the problem posed by the task" (p. 242). Open tasks are defined as those which have "no single correct answer that learners must identify." Willis and Willis (2007) also provide a succinct description: "A closed task is one where there is a "correct" answer, for examples in a "Spot-the-difference" task where there are five differences to be found. An open task is where the outcome is unpredictable – where learners are free to decide what they want" (p. 156). In terms of the effect of each task type on learner output, Long (2014) writes that closed tasks may promote more fluency, and open tasks may promote more accurate and complex output.

2.3.7 Task repetition

Having introduced the two primary models for assessing task-type and L2 acquisition, and the task type used in this study, this section focuses on another crucial element: task repetition. As seen in Skehan's meta-analysis (2016), task repetition is attributed to having a potentially significant impact on output complexity, where learners may produce more complex output during task repetitions.

Task repetition can involve the repetition of the same task, or the same task in an altered form (Bygate & Samuda, 2005). From a psycholinguistic perspective to SLA,

research has shown that task repetition can positively affect L2 development by promoting faster access to language items (Larsen-Freeman, 2012), increase access to language components (Ahmadian, 2011), and can free up cognitive capacity to focus on forms (Bygate, 1996; Fukuta, 2016). This can result in an improvement in learner output in terms of complexity, accuracy or fluency. However, as Skehan's model predicts, only two of the three skills are reported as being improved. Learners in Bygate's (1996) study showing an improvement in complexity and small gains in accuracy, yet in a subsequent study (2001), an improvement in fluency and complexity was recorded, but not for accuracy. He argues that during the initial task, learners are more concerned with the heuristic planning of content, whereas on a second attempt, learners, who are now familiar with the task contents, can apply more of their cognitive capacity to linguistic formulation. Additionally, in a study by Hawkes (2011), learners seemed to focus more on form during their second encounter with a task, which supports the findings of Bygate (1996) and Fukuta (2016).

A recent study by Sample and Michel (2014) explored how learner oral task performance during a spot-the-difference task improved over multiple repetitions of the same task. In other words, instead of looking at a single repetition, learners in this study completed the same task three times. Upon completion of the first repetition, Skehan's trade-off hypothesis appeared to be accurate; learners improved their performance in terms of fluency at the expense of complexity and accuracy. However, during the second repetition, trade-off effects disappeared, leading them to conclude that with growing task-familiarity learners are able to focus their attention on all three CAF dimensions simultaneously.

Another issue with task repetition is Skehan's (1991) Resultative Hypothesis. The claim made is that the mastery of tasks can affect learners' motivational attitudes. More specifically, learners who do well at a task persevere and maintain or increase motivation. This has relevance to the current study in that the language goals are the same for both the VW and FTF tasks in a task pair. Therefore, the first task of a task pair may be perceived as more difficult than the second task due to their unfamiliarity with the task demands and lexical content, producing a negative effect on the student motivation. In turn, the second task may be perceived as less difficult, and induce a positive effect on their motivation. In summary, a counterbalanced approach to completing tasks may help mitigate the effects of task repetition on participants' output and perceptions of task difficulty.

2.3.8 Section summary

Based on the literature of task conditions and output complexity, accuracy and fluency several considerations must be made. Firstly, task design is only one variable in determining output. Learner characteristics such as familiarity with a task, as well as affective variables such as the age or sex of their interlocutor may also affect performance. Contextual demands are also considered a source of task complexity, where conducting a task face-to-face may be considered less demanding than tasks mediated by technology. Additionally, research design considerations should ensure that task complexity conditions are kept as close as possible between tasks in a task-pair as they have the potential for influencing output the most. Finally, as task repetition has the potential to influence output, motivational characteristics and perceptions of task difficulty, the sequence that participants complete tasks should be counterbalanced.

2.4 VIRTUAL WORLDS AND VIDEO GAMES: A DEFINITION

This dissertation explores the effect of cognitive task complexity and modality on low-level learners' oral task performance. This is achieved by requiring learners to complete tasks via two different modes of communication: online (within a VW) and offline (face-to-face). This section defines the concept of "virtual world," as used in this dissertation and wider CALL literature. The genesis of VW technology is described, along with how VWs have been utilised in second language learning contexts. The subsequent sections also focus on the potential benefits and hindrances of computer-mediated communication (CMC) as well as introducing several theoretical frameworks for teaching languages with digital games, where the conceptual differences between VWs and digital games are made salient.

Virtual worlds have been in existence since the early 1980s, the first of such known as multi-user dungeons (MUDs). MUDs still exist today and are defined loosely as text-based adventure games allowing users to engage in real-time communication, develop their characters and role-play (Bartle, 2003). The first MUD was an extension to the popular single player game – *Zork* developed by Roy Trubshaw in 1978 called *MUD1*. It was the first Internet-based multiplayer video game when Essex University connected its pre-existing local area network to the ARPANET in 1980. Simple text-based commands such as WALK, OPEN, CLOSE, and PICKUP allowed individual users to manipulate virtual objects,

navigate the virtual world, and interact with other players and non-player characters (NPCs). MUDs typically feature fantasy worlds populated with dungeons to explore, fictional monsters to fight, and a variety of classes from which the player may choose (see Figure 1). In many cases, these elements reflect the play style of the predecessor to MUDs: the *Dungeons and Dragons* series of tabletop roleplaying games (Gygax & Arneson, 1974).

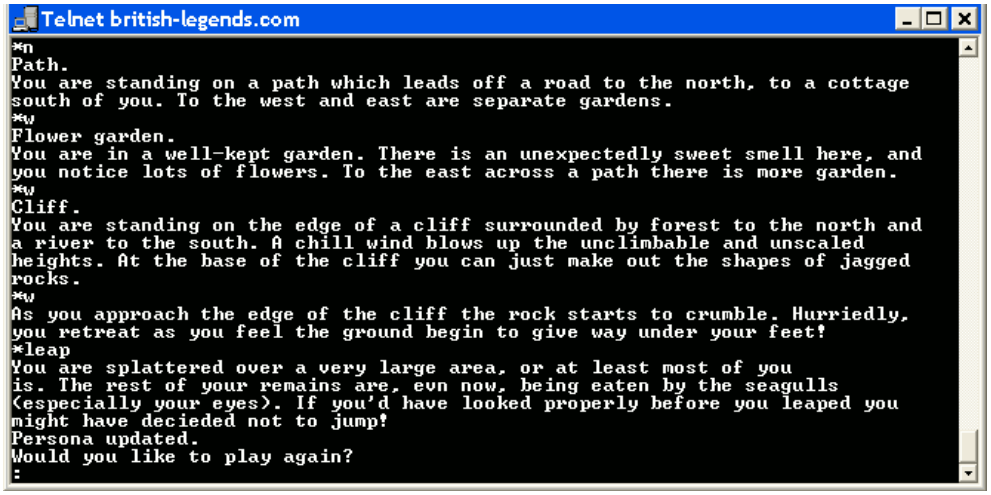


Figure 1: A screenshot of MUD1 (Wilks, n.d.)

With advances in computer programming and the worldwide web, MUDs evolved into object-oriented MUDs (known as MOOs) which allow players to not only interact with objects and users but also create their own content within the world (Hayes & Holmevik, 2001), resulting in more advanced interactivity and game-playing depth. Compared to MUDs then, MOOs represent a read- and write-enabled interface, with content produced by players via an object-oriented programming language. One of the implications of this is that players could expand game content indefinitely. The first MOO appeared in 1990 named *AlphaMOO*. However the improved version *LambdaMOO* (released in 1991) was a more popular version and still available to play today. MOOs have since been adopted in educational contexts for distance and blended learning, collaboration, and teaching object-oriented concepts (Bartle, 2003; Peterson, 2009).

Coupled again with technological advances including affordable Internet access, high-spec home computers and video games consoles capacity to display more and more elaborate graphical components, MUDs represent the precursor to massively multiplayer online role-playing games (MMORPGs or MMOs). Contemporary MMOs can thus be considered a natural evolution of early MUDs and MOOs, showing their ancestral traits such

as player-to-player interactions, class selection, a levelling system, experience points, and other, similar game mechanics. The popularity of the MMO genre has led to the development of a plethora of themed worlds, with varying degrees of interactivity and differing game mechanics.

World of Warcraft is an MMORPG set in an archetypal fantasy setting where players may choose to play on either side of two opposed factions loosely representing the humans and *Orcs* of the original *Warcraft* series (Blizzard, 2004). Play involves defeating increasingly strong non-player characters (NPCs) which are typically humanoids or monsters and levelling up your character's skills and abilities. Players can reach the highest level of the game (currently level 110) via several fairly linear routes, but typically either complete quests or defeat other players in player-versus-player combat. Upon reaching the highest level players gain access to some of the game's fiercest enemies, which can only be defeated by carefully choreographed groups of players numbering up to 40. There are four different game modes to choose from, designed to suit individuals' play style preference and goals (Table 4).

Table 4: Game modes in World of Warcraft.

Game mode (server type)	Playstyle
Normal	A typical player versus environment (PvE) where play is focused on defeating high-level monsters and quests. Players are unable to attack each other in the open world freely. As such, player-versus-player (PvP) combat must be consensual.
PvP	As well as the above quest and exploratory elements, PvP servers allow for player-versus-player combat at all times, and in all areas of the world.
Roleplay (RP)	RP servers are similar to normal servers, but there is a strict policy of roleplaying the character in the game.
RP-PvP	This mode is the same as the PvP server above but integrated with the rules of an RP server regarding playing the game in character.

In contrast, *Second Life* is a virtual environment that first appeared in 2003 from Linden Labs. It is often categorised as a multi-user virtual environment (MUVE) and differs from MMOs drastically in that it dispenses with any set gameplay, progression models, and

narrative. Instead, it focuses on social networking and world-building such as that typically found in MOOs (Peachy, Gillen, Livingstone, & Smith-Robbins, 2010). While it is true that MUVEs existed before modern social-networking sites (SNSs) such as *Facebook* and *Twitter*, the development and evolution of MUVEs has subsequently been influenced by the technology of SNSs. For example, the creation of a user profile, public and private status updates, and community and group participation are common elements of SNSs that now appear to some extent in MUVEs. Using a three-dimensional modelling tool (see Figure 2), players can create virtual objects in Second Life and additionally, using a specially developed programming language, add interactivity to their creations.

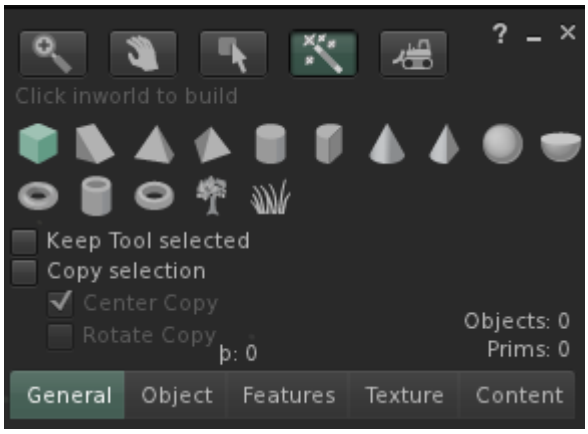


Figure 2: Second life object creation tool (Linden, 2018).

Although MMORPGs and MUVEs like Second Life differ significantly, they share several features which allow them to be classified together under the umbrella term "virtual world." According to Smart, Cascio and Paffendof (2007), these features include:

- In-game environment persistence,
- An area to host simultaneous connections from multiple users,
- The 3D representation of a player in the form of a customizable avatar,
- Interactions between users and objects in a 3D virtual world,
- Real-time feedback to players of any such interactions,
- Physical and temporal similarities to the real world, to provide a sense of existence within the ecosphere of the game.

The majority of MMOs and other virtual worlds feature a way to communicate with

other players, typically via text-chat, but more recently games feature in-game voice chat systems as internet speeds continue to increase.

This section has briefly outlined the emergence of virtual worlds, where the main evolution is shown in Figure 3. The technology employed in this study may be considered a MUVE of sorts. However, more detail on how and why the environment was chosen is provided in the methodology chapter below.

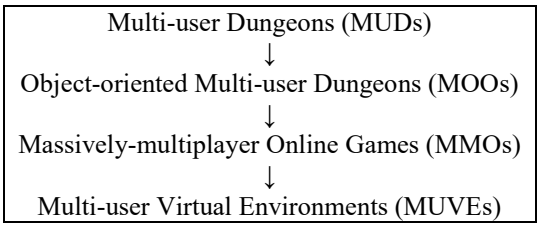


Figure 3: Simple evolution of virtual worlds.

2.5 LANGUAGE LEARNING WITH DIGITAL GAMES AND VIRTUAL ENVIRONMENTS

The use of virtual environments in language learning contexts is closely related to the field of "Digital game-based language learning" (henceforth: DGBLL) which has grown to become a significant sub-category of CALL research over the last decade (Cornillie, Thorne & Desmet, 2012). As games and VWs become more immersive, their potential benefits for language learning have become more pronounced, resulting in increased interest regarding their effective implementation in language learning contexts. The rationale for adopting virtual worlds in instructed SLA environments is often due to the apparent *authenticity* of experiences that these domains provide, as well as their rich *affordances* for language learning. It is worth taking time here to unpack both of these words (that is, "authenticity" and "affordance").

2.5.1 Task authenticity and virtual worlds

The importance of authenticity is stressed in communicative language teaching and communicative pedagogies because of the assumed ease with which learners will be able to transfer skills from an "authentic" classroom to real-life language use contexts. That is, if the tasks learners complete in the classroom are similar (i.e. authentic) to the tasks they are required to do outside of the classroom, or that members of the target culture are likely to engage in, transfer may be more easily achieved (Kramsch & Thorne, 2002). In terms of authenticity in CALL, Buendgens-Kosten's (2013) introduced three domains of authenticity:

linguistic, cultural, and functional authenticity.

Linguistic authenticity refers to the level of authenticity of materials supplied to the learner. For example: if teaching materials feature 'normal' or 'natural' language, produced not for language learners but native speakers, these materials could be considered linguistically authentic or 'real.' In CALL, there are multiple avenues for connecting learners to linguistically authentic language. Some examples include the internet, social media, and, relevant to the present study, the interactions that learners may have with non-player characters and other players (often native speakers) as part of digital gameplay.

The *cultural* authenticity of teaching materials refers to the level of connection materials have to the target culture similar to the concept of "realia." Therefore, cultural authenticity does not specifically relate to the language used in materials but the origin of the materials and cultural-closeness of content. For instance, a beginner class of Japanese learners may encounter flashcards with words for *uwabaki* (indoor shoes), *gochisousama* (a phrase said after eating), and *osakini shitsurei simasu* (a phrase said before leaving work) which reflect the cultural norms of the Japanese. Such an activity presents artefacts of the Japanese culture while being a language learning activity at the same time. In CALL, cultural authenticity is often complexly intertwined with linguistic authenticity in video sharing sites like YouTube and social media where people (generally) represent a particular culture associated with the target language. Additionally, cultural authenticity may be connected not to a particular geographical location, but an affinity group (Gee, 2005) around a popular cultural item or pastime. For instance, affinity groups which have their own culture, vernacular, and codes of conduct appear around games (eSports, MMOs, console-specific, genre-specific), books and movies (fan-fiction), and so on.

Finally, *functional* authenticity is considered the "ordinary practices of the culture" (Brown, Collins and Duguid, 1989 p.34). Buendgens-Kosten (2013) outlines how functional authenticity relates to linguistic and cultural authenticity with an example of a local weather forecast from the target culture used in the classroom. Whilst the weather forecast is culturally authentic (as it is an original source) and the language used by the broadcasters' is linguistically authentic, if it is only used as a linguistic resource (to mine for words related to weather for example) then the function or practical application of the forecast is low in

functional authenticity: it is not used to plan a trip, decide what to wear that day or whether one should take an umbrella to work or not. This episode also emphasizes the functional authenticity of the *target* culture, not the culture of the learner, which is also of consideration. In other words, is the learning task related to the particular functional needs of the learner?

In relation to VWs, *linguistic* authenticity may be increased by using examples in the pre-task phase of native speakers carrying out the tasks designed for the learners. *Cultural* authenticity may be produced in VWs by immersing learners in cultural experiences such as in work by Shih (2015) who had students take a "virtual walk" through London with the use of Google Street View technology coupled with VR headsets. Additionally, Sykes (2014) used photos and other realia collected from the target culture as props in her custom VW to increase immersion and thus authenticity. Alternatively, the VW domain itself could be considered an artefact belonging to the target culture, developed within the realm of the target language, and thus culturally authentic. For instance, the game *Shenmue* (Sega AM2, 1999) is set in a specific suburb of Japan and features a staggering amount of cultural detail to both the place and time that the fictional story is set.

Finally, f*unctional* authenticity also relates to the immersive nature of virtual environments. The argument has been made that if the virtual landscape replicates the domain of L2 usage more closely than that of traditional classrooms, learners may perceive tasks as being more functionally authentic (Park, 2018; Shih, 2015; Sykes, 2014). Indeed, this is one reason why foreign language educators were early adopters of the technology: to provide learners with "telepresence" or the sense that they are in a "real" space with their interlocutor (Blyth, 2018).

The authenticity of language use as part of digital game play or social activity within a VW may seem far removed from real-life needs. However, gaming allows for meaningful opportunities to use the L2 in situated contexts (Gee, 2004). Additionally, if the students self-identify as gamers, the function of learning a language as part of gameplay or using the target language to conduct gameplay may be considered a suitable and relevant activity, thus increasing perceived functional authenticity. Related to this point, Blume (2018) writes that digital games allow students to be their "authentic selves" where games may "reflect their core values and interests" (p.25).

In summary, VWs have been considered a beneficial domain for technology-mediated TBLT as they provide learners with more meaningful, *authentic* experiences than other online learning spaces due to their innate *affordances* (Sykes, 2014). The word "affordance" however, requires additional explanation in order to make its meaning and relation to virtual environments more salient.

2.5.2 Affordances of virtual worlds

Affordance as a term within the literature on language learning is most strongly associated with the work of Van Lier (2000, 2004, 2010). Van Lier takes an ecological approach to language learning, inspired by Gibson's book on the ecology of visual perception (1979). Ecological, in this case, refers to the broad spectrum of academic, professional, and pedagogical "work" of teaching and learning (2010, p.3), thus incorporates the actions of both teachers and learners in the learning process.

An *affordance* is considered the relationship between an "organism" and its environment that either allows for or inhibits an action (2004). That is, an environment is said to provide a "semiotic budget" from which action may emerge (2010). One may consider this not as input, but as a type of potential energy for further, meaningful action. What exactly becomes an affordance is based on what the "organism" does or wants to do within the environment. As a concrete example, instead of having access to a hammer, a human may choose a heavy rock which happens to be in the environment to break open a coconut. The rock, therefore, affording the action of breaking open a coconut but not being the cause or impetus of the action. For instance, if the human did not have access to the coconut but was tired from engaging in strenuous activity, the rock could afford the action of sitting down and resting. The relationship between the actor and environment thus providing the impetus for a specific action.

In terms of language acquisition, a learner's level of engagement and activity within the environment is said to affect whether a linguistic affordance (which may be utilised for linguistic action) is noticed or not (Darhower, 2008). For example, a learner could be exposed to the L2 and not attend to it at all, could register it and move on without questioning it, or could actively and critically process the input for comprehension. In this way, the affordances for language learning of a specific environment involves the interplay between

a learner's capabilities and the opportunities for learning provided by that environment.

VWs are considered to have a rich semiotic budget for language learning (Newgarden, Zheng, & Liu, 2015), with affordances for safe learning and languaging (Rama et al., 2012), goal-directed action (Peterson, 2012), and non-verbal avatar-mediated communication and comprehension (Zheng, 2016). Additionally, Darhower (2002, 2008, 2013) explored the linguistic affordances of telecollaborative projects via CMC. His study found that L2-speaking interlocutors afforded a large volume of feedback on learners utterances, thus significantly increase the semiotic budget of the learning environment.

Having introduced the core technology utilised in this study, along with the two key concepts that underpin the rationale for its adoption, the literature review now introduces the history of games-based studies in CALL.

2.5.3 The genesis of DGBLL in CALL

Language teachers and researchers have been exploring the connection between games and language learning since the first videogames were developed. One of the earliest pieces of software designed specifically for language learning is an interactive multimedia game titled *A la Rencontre de Philippe* (Furstenberg and Malone, 1993) which was considered "a major step forward in the teaching of French" (Gray, 1992). The game is similar to a modern adventure role-playing game where the player must help the protagonist Phillipe find a new apartment after being told to move out by his girlfriend. The social drama provides an immersive context for the player, and the ensuing story features both text and oral input, promoting the development of several important skills for language learners. This type of game is typical of the era and may be described as "game as *tutor"* (Thorne & Reinhardt, 2016).

Thorne and Reinhardt (op cit.) differentiate between the ways games have been utilised in CALL programs over the last two decades. This ranges from the early tutor-like programs that offer one-way, multiple choice, interactive activities to learners, to the present day where we see games as medium or method of authentic L2 acquisition in cases like MMO participation. As games have evolved over the last few decades, researchers have been developing frameworks to guide pedagogical implementation that make use of the specific affordances of games (including virtual worlds) for language learning (e.g. Higgins 1983;

Bax 2003; Sykes & Reinhardt, 2013). However, there is currently no single theoretical or pedagogical approach to conducting DGBLL. Trends in DGBLL research include learner perceptions of their own learning gains (Wang, Petrina, & Feng 2017; Peterson, 2012), assessing the affordances of games for language learning (Rama, Black, van Es & Warschauer, 2012; Reinhardt & Sykes, 2012), and the applicability of games for language learning from a policy-maker or educator perspective (Franciosi, 2015; Hourdequin, York & deHaan, 2017).

Hung et al. (2018) conducted a recent meta-analysis of DGBLL research, highlighting that the "majority of DGBLL studies featured positive outcomes regarding student learning, with the most frequently reported ones related to affective or psychological states" (p.89). This focus on affective conditions may be considered a cause for concern. Indeed, Peterson points out that it is the design of appropriate learning tasks and content that make use of the affordances of technology that is most likely to maximise learning opportunities (Peterson 2010a). Additionally, Peterson (2013) also emphasises that there is a heavy focus on vocabulary acquisition studies within the DGBLL literature, and therefore a need for studies which explore other areas of language learning.

The current study is thus positioned in just such a domain: to provide empirical evidence for the language learning potential of VWs in terms of how online communication within such environments may affect learners' oral performance. Subsequently, the next section reviews the various existing DGBLL frameworks through the lense of Reinhardt and Sykes' classification (2014) as a way to illustrate how the current study can be considered a *game-based* approach to DBGLL. After introducing the various frameworks, this section reviews relevant literature relating to three key areas: (1) the general benefits of VWs for language learning, (2) studies that are concerned explicitly with the application of TBLT in VWs, and (3) studies that compare modality with a focus on CMC. Upon reviewing the literature, the subsequent section identifies a need for studies that explore oral SCMC and task-development for VWs.

2.5.3.1 Research frameworks for DGBLL

There have been numerous attempts to integrate games into second language teaching and learning contexts as seen in papers by Gaudart (1999), Baierschmidt (2013), and Hastings

(2014). Reinhardt and Sykes (2014) provide a classification system for DGBLL which separate the approaches based on the type of game being used. For an overview of the three models, see Table 5. By referencing their classifications, the current study can be considered *game-based*, distinct from *game-enhanced* and *game-informed* approaches. Following is a critical evaluation of each approach.

Table 5: Digital games and language learning research practices (adapted from Reinhardt and Sykes, 2014).

Model	Features	Research questions
Game-enhanced	Use of vernacular, off-the-shelf games (i.e., games designed for entertainment purposes)	How can vernacular games be pedagogically-mediated for L2 learning and teaching?
Game-based	Use of educational or learning-purposed games (i.e., synthetic immersive environments)	How can game-based environments be designed to incorporate and complement L2 pedagogical uses?
Game-informed	Game and play principles applied in digital and non-digital contexts outside the confines of what one might typically consider a game	How can insights from the study of games and play inform our understanding of L2 teaching and the design of all L2 learning environments?

2.5.3.1.1 Game-enhanced research

The use of commercial, over-the-shelf (COTS) games, designed primarily for entertainment, can be beneficial for language learning due to their unfiltered, authentic target language content. Compared to specifically-designed, language learning games or environments, COTS games do not provide tutorial content explicitly aimed at developing one's language-learning skills. Instead, they can provide learners with a rich world of text, and authentic native-speaker communication (Cornillie, Thorne, & Desmet, 2012).

Due to the lack of support for language learners when playing COTS games, game-enhanced research often centres on educators creating scaffolding frameworks and pedagogies to help learners become accustomed to the game and exploit the language learning potential of game content as much as possible. One example of this is seen in York and deHaan, (2018) where a specific pedagogical intervention for the implementation of games into a TBLT approach to SLA was created (see also York, deHaan & Hourdequin, 2019).

Miller and Hegelheiner (2006) investigated how the COTS game "The Sims" could be used to teach both vocabulary and grammatical items to adult learners. Their study design implemented several supplementary materials that were available to an experimental group but not the control group. Findings showed that the experimental group retained more

vocabulary than the control group, thus suggesting the importance of pedagogical considerations and support for game-enhanced contexts. In other words, by merely playing vernacular games without such support, learners may not learn the content.

One of the most rigorously considered models for game-enhanced language learning is the Explore, Examine and Extend (henceforth EEE) sequence, created by Reinhardt and Sykes (2012). However, unfortunately, as of writing this, it appears that the model only exists theoretically, with no empirical studies on its implementation appearing in the literature.

The aim of Reinhardt and Sykes' model is for other practitioners to be able to implement video games in their own teaching contexts successfully. Table 6 provides an overview of the model.

Table 6: EEE sequence overview.

Framework phase	Suggested activities
Explore	Basic playing of the game.
	Observing the game being played.
	Noticing particular items or collecting discourses with guidance.
Examine	Playing the game with a more intensive focus on discovery.
	Completing analysis activities on discourses targeted to meet the specific linguistic, pragmatic, or sociocultural objectives of the lesson.
Extend	Active and reflective creation of new discourses with or through the game.
	Participation in attendant discourses.

Initially, learners *explore* the game space. Activities at this phase require learners to learn about and collect discourses from within the game such as the language used, overarching stories, or other narrative elements. During this exploratory phase then, learners become familiar with the game rules, and context of the game. Following is the *examine* phase where learners play more critically. Attention may be focused on linguistic items, cultural artefacts, or winning strategies. Upon analysing the game, the authors promote the idea of getting learners to experience analysed discourses in more real-world contexts that promote interaction:

> Students experience the discourses with new contextualised understanding in the game, through the game with other players, or around the game in a structured activity designed to promote interaction. Learning is facilitated by using the discourses in a meaningful, experiential way after having analysed them. (Reinhardt & Sykes, 2012, p. 6)

Reinhardt and Sykes recognise that video games may not provide an adequate context for promoting interaction between learners, and thus they prescribe a specific phase in their model for post-play social interaction. The structured activity designed to promote interaction mentioned above describes the idea of a focused task from a TBLT perspective (see Ellis, 2003).

The *extend* phase is designed to get learners participating in discourse circles in and around the game by prompting them to become involved in discussion board posts, creating fan fiction[1], producing machinima,[2] or reviewing the game. Studies concerned with fanfiction and English language learning have been growing in the last 10 years literature (Black, 2005, 2006, 2009a, 2009b; Thorne, Fischer & Lu, 2012; Gee & Hayes, 2012), yet from my own experience, only a certain, small proportion of L1 players actually participate in such activities. As such, participation in fanfiction may be considered an advanced form of participation suitable for motivated, higher-level L2 learners (i.e. learners that differ significantly to those in the current context). In addition, as only real "fans" of the game participate in such activities, it may not be an enticing or motivating activity for most L2 learners. Indeed, only a few studies exist regarding the development of fanfiction as a pedagogical tool, and these are in an advanced language learner context (see Sauro, 2014; Sauro & Sundmark, 2016).

In summary, this model, despite the lack of empirical studies, represents a robust, TBLT-inspired approach to the use of video games as a teaching tool in language learning contexts. The framework's socially informed underpinnings are evident in the concept of

---

[1] Fiction created by fans of games using the specific game universe as inspiration for an original story (see Black, 2006 for an overview).

[2] Short movies made using in-game content (**machin**e + cin**ema**).

interacting with games (ideational interactions), through and around games (interpersonal interactions) and about games (textual interactions). Therefore, it builds on the work of Miller and Hegelheimer (2006) in creating a more robust support framework for the use of vernacular games in classroom-based language learning environments. Such game-enhanced approaches to SLA, however, are not the focus of the current study. This study is firmly situated in what Sykes and Reinhardt consider *game-based* research, which is introduced in the next section.

2.5.3.1.2 Game-based research

Game-based research is concerned with the development and implementation of games designed for specific educational purposes. This type of research therefore differs significantly from the previously introduced "game-enhanced" research. From a language learning perspective then, games may be developed to encourage players to both encounter and interact with a foreign language. Traditional game-based tuition sees learners interact with a computer and receive feedback on their inputs. An example can be seen in Figure 6, which represents a simple listen-and-click game to learn the Japanese writing system: hiragana.

Figure 4: An example of a simple educational game for learning Japanese hiragana (Gibson, n.d.)

One criticism of the products, like this, created as part of game-based research projects is that the focus on learning outcomes is too explicit and may be considered "sugar-coated" with extraneous game mechanics in order to motivate students to continue playing. In other words, compared to COTS games which are first created from an entertainment perspective, games that are created with a learning outcome *first* often try to obfuscate the learning content with the (often arbitrary) addition of game-like systems. This has been referred to as "chocolate-covered broccoli" (Bruckman 1999) where the gaming element of the product is used as a separate reward upon completion of the learning content. The potential mismatch between educator goals and student motivations for playing is a critical factor when designing such simulations. It has been shown that individuals with pre-existing gaming experience prefer not to play educational games due to their preconception of what a video game is, and the inability of educational games to meet their expectations (Chik, 2014).

In response to the negative reception of educational games, the use of synthetic environments for language learning has become a more prevalent theme in the research literature. These 3D environments add further avenues for interaction such as learner-to-learner or learner-to-space modes. For example, Henderson, Huang, Grant and Henderson (2012) created an interactive virtual environment as a tool for learning Chinese. Additionally, Peterson (2012) investigated the sociocultural affordances of a virtual environment for language learning with Japanese EFL learners. Cornillie, Clarebout and Desmet (2012) created a virtual learning space to investigate learner perceptions on the use of corrective feedback. Finally, Wang, Petrina and Feng (2017) developed a Virtual Immersive Language Learning and Gaming Environment measuring participants level of immersion based on the addition of certain non-player characters (NPCs). The language learning potential and benefits of such environments are numerous and are covered in more detail in the section below (Section 2.5.4).

2.5.3.1.3 Game-informed research

Game-informed research, as coined by Sykes and Reinhardt (2013) and Reinhardt and Sykes (2014), is concerned with the application of the highly motivating mechanics found in digital games to real-life situations. This type of research is widely known as *gamification* (Kapp, 2012). The term first appeared in 2002 (Paharia, 2010) and was adopted

more predominantly from 2010 (Deterding, Dixon, Khaled, & Nacke, 2011). It has been described as "game-thinking and game mechanics to solve problems and engage audiences" (Zichermann & Cunningham, 2011), or a way "to make non-game material more engaging" (Takahashi, 2010). Examples of gamified educational contexts include Sheldon's (2011) "Multiplayer Classroom" concept which aimed to apply the game mechanics often found in MMOs to evaluate his students. In language learning contexts, York (2012) emulated Sheldon's original concept specifically for use in an EFL classroom environment. Lombardi (2015) also created a gamified English class where students progress through a points-based system of in-class tasks and homework activities in order to attain a particular grade.

Gamification is a wildly popular concept not only language learning contexts but in wider educational fields, where it is often espoused as a key way of increasing student engagement and satisfaction in their learning with minimal effort required of teachers (Urh, Vukovic, Jereb, & Pintar, 2015). A magic bullet of sorts. However, there are a considerable number of critics of the approach who claim that it is not a new concept or even that effective (Seaborn & Fels, 2015). Indeed, Nicholson (2015) writes that "meaningful" gamification requires a great deal of effort from educators to implement successfully.

Having introduced Sykes and Reinhardt's three categories for conducting DGBLL research, the next section focuses on the language learning potential of digital environments. Both game-enhanced (use of COTS games) and game-based (the development of games for learning) models are considered with a focus on MMOs and social worlds in particular.

2.5.4 VWs as sites for language learning

There are a growing number of studies that explore the use of VWs (often referred to as "games") for SLA. However, the field is somewhat split between game-enhanced and game-based approaches to DGBLL. The two overarching categories of VW currently receiving attention from CALL scholars are MMOs and social worlds. MMOs represent commercial off the shelf games that are developed by a specific company for entertainment purposes. As such these environments are often studied in terms of their affordances for extracurricular language learning, where both affective and cognitive affordances have been explored (see Peterson, 2016 for a recent meta-analysis). Social worlds, as introduced above, are venues which allow the instructor to develop language learning activities themselves.

Such environments are therefore often linked to studies that explore how SLA methodology may be applied in these domains, where best practices are theorized (Jauregi et al., 2011; Gánem-Gutiérrez, 2014). This section takes a closer look at studies concerned with social interactions in online worlds and the possible benefits the domains hold for language learning. Upon outlining general affordances and benefits, I introduce those studies that explore the adaptation of TBLT for use in VWs and then the benefits of CMC with a focus on those studies which compare modality.

2.5.4.1 Overall benefits of learning in immersive virtual environments

Although MMOs are different to the type of VW employed in the current study, their language learning affordances may be applicable and thus deserve attention. Additionally, they represent the most widely studied VW, and so there is a wealth of research to call upon. Peterson (2016) provides a meta-analysis of 10 studies, which are concerned with how MMOs may be utilised for successful language learning. Studies were selected for inclusion if research design drew on cognitive or sociocultural accounts of SLA. Of the ten studies, Peterson highlighted the language learning benefits of MMOs from both research paradigms.

Cognitively, undertaking activities in MMOs has been shown to help lower the affective filter in several ways. First, players may *co-act* with a customizable avatar through which interactions with other players take place (Zheng & Newgarden, 2012; Newgarden & Zheng, 2016). Thus, avatars provide a mediating tool onto which players may project themselves and communicate with others. Secondly, online environments are considered safe spaces for learners to experiment without risk, a feature that has been linked to the improvement of learners' willingness to communicate (Reinders & Wattana, 2011; 2014; 2015). Indeed, Blume (2018) posits that the affordances of games for providing a safe space for hypothesis testing and making errors is far higher than that of classrooms or offline interactions which are often "fraught with communicative pressures" (p. 25).

The implications of this for the current study are that learners may produce more errors when completing tasks in a VW. Thus a decrease in accuracy may be predicted. A further complication is that the VW tasks are predicted to be of higher task complexity than the FTF tasks, which according to the claims of CH, also suggests that accuracy will be reduced when the (+/-) here-and-now task condition is manipulated to be more complex

(Robinson, 1995; Rahimpour, 1999). However, affectively, learners may prefer to conduct tasks in the VW due to the affective affordances. Online worlds are also presented as TL-speaking environments where learners are exposed to large volumes of rich input both from the game itself (Thorne, Fischer, & Lu, 2012) and other players (Zhao & Lai, 2009). Learners are also required to produce TL output, making them a motivating domain for promoting autonomous learning (Lee & Pass, 2014).

From a social-informed perspective, in-game quests may require collaboration with other players in order to be completed successfully, providing learners with opportunities to collaborate in the TL with other, often more-experienced players. Such interactions may be likened to expert-novice interactions in zones of proximal development (Rama, Black, Van Es & Warschauer, 2012). MMOs also feature game-related communities of practice for learners to socialise with others in what Gee (2004) calls *affinity spaces* (see Thorne & Black, 2008; Lee & Gerber, 2013). Peterson (2012) analysed the discourse of four intermediate EFL learners as they played an MMO together and found that the environment provided the impetus for collaborative social interaction such as the use of positive politeness. DuQuette (2013) also notes that the complexity of the environment may catalyse beginner-expert interactions. This is seen in exchanges between novice users asking for assistance from more experienced players in order to navigate the virtual environment effectively.

The specific affordances of MMOs for language learning may be exploited in *game-based* learning environments also. For instance, in the current study, where there is no access to native speakers, the development of an affinity space may be difficult, but other affordances such as avatar-based interaction and tasks that require collaboration between learners may be created. If designed with a focus on learners' language development, such tasks may prove more effective than those tasks found in MMOs. It is to this point that the literature review now turns where one particular affordance of MMOs that informs the current study is introduced in more detail: Quests.

2.5.4.2 In-game quests

Zheng (2012) identified and distinguished between the different coordination and language activities of English language learners during an episode of World of Warcraft gameplay. One finding was that the environment produced a large volume of communicative

activities between learners that were unlikely to be matched by what occurs in a classroom environment. For example, in a single gaming session, a total of 13 communicative activities sub-types were identified such as offering help, seeking help, reporting on actions, locating, apologising, and utterances on how to use the technology.

From this, they make suggestions for the design of learning environments based on affordances for co-action and rich communicative activities. Specifically, they note that the concept of "questing" as an important affordance for L2 acquisition. Quests are activities given to players that teach them about the culture of the gaming environment. Such quests can be considered "tasks" from a TBLT perspective, as they are meaning-based activities with concrete goals. However, quests have also been described as "incentivised chores" (Landwehr, Diesner, & Carley, 2009) due to their often repetitive nature, and a low variance in available quest types.

In the case of World of Warcraft, the most popular MMO currently, there are thousands of quests, but analysis reveals only a few common quest types (see Table 7) (WowWiki, 2016). If these few, pre-determined quests make up the bulk of MMOs tasks, and the literature reports that MMOs are a successful and important domain for language learning in the 21st century, I argue that with teacher-designed, language-learning specific tasks in a VW, there are potentially many more affordances for successful, directed language learning with an emphasis on speaking skills. This is one of the main reasons for rejecting MMOs in favour of VWs for the current study. An example of a teacher-defined task (or "quest") to enhance spoken output is to separate information regarding quest objectives between interlocutors, forcing them to communicate. Such "gap-like" considerations are not found in MMOs, where all players have access to the same information from the start of a task.

Table 7: Task types in World of Warcraft

Quest type	Typical objectives
Gather	Gather a number of items and return to the quest giver.
Kill	Kill a number of creatures and return to the quest giver.
Deliver	Deliver an item for a quest giver to another NPC.
Loot	Kill a number of creatures, collecting a specific number of items upon killing them, and then return to the quest giver.
Create	Use a profession to make an item and return with it to the quest giver
Escort	Escort an NPC from one place to another.
Find and speak	Find (and speak) to an NPC.
Befriend	Build a level of positive reputation with a specific NPC faction.
Explore	Explore a particular region or area and return to the quest giver.

In summary, Zheng concludes that providing learners with concrete goals in the form of "quests" is considered a useful way of promoting L2 acquisition. However, I argue that the quests found in MMOs are too limiting in terms of the interaction they afford and propose that VWs may offer better opportunities for instructors to create appropriate language learning tasks.

2.6 TBLT IN VIRTUAL ENVIRONMENTS

Studies introduced in the above section outlined the general affordances of games as environments for language learning with a focus on MMOs. In this section, I shift to focus more specifically on *game-based* studies (the design and implementation of games for educational purposes), and in particular those that explore the application of TBLT in virtual environments.

Jauregi et al. (2011) developed a set of design principles for interaction tasks in virtual worlds. Their specific aim was maximising authentic social interaction and intercultural awareness as part of tasks while utilising as much of the virtual world's affordances as possible. They created a detailed list of task design principles for virtual worlds; however, the tasks they eventually chose for their learners do not seem to reflect their strict design criteria. For instance, two of the four tasks did not require the use of a VW and could have

been completed with simple videoconferencing software. Additionally, their results suggested that participants spoke more when completing these tasks that did not explicitly require the use of the VW. One caveat, however, is that the extent that learner output differed between the tasks they completed was not explored in any empirical way due to it not a major focus of the study.

Assumptions as to why the VW-specific tasks did not promote as much communication between interactants are that 1) VW tasks were not designed rigorously enough, or that 2) requiring learners to interact with the VW actually hindered production. The authors conclude that their design principles could be improved upon, "Task design principles have to be further specified for 3D virtual world settings, focusing on enhancing rich oral interaction to be necessary for task completion, while exploiting at the same time the exploratory, functional, and gaming possibilities of Second Life as much as possible" (Jauregi et al., 2011 p.97). The aim of the current study is to investigate this notion in more detail. Tasks in this study were designed to require learners to interact with and navigate the virtual terrain as an essential element of task completion. With the creation of tasks that take advantage of the affordances of VWs in this way, it may be possible to promote such "rich oral interaction."

DuQuette and Hann (2010) also explored learner interaction in Second Life via custom (researcher-created) information gap tasks. The first task employed was an information gap activity based on the concept of giving and receiving directions. Learners guided each other through the VW based on a predefined route. The second task was a two-way information gap activity, where information was shared between the participants. This task required learners to give instructions to their partner in order to re-arrange furniture based on a predetermined arrangement.

It should be noted that this study is one of very few that requires the learners to manipulate the environment (i.e. rearrange "physical" objects in a virtual room), thus making more use of the VW's affordances than other studies highlighted so far. As a side note, one reason that researchers may be reluctant to get learners to interact with virtual environments in this way could be due to the inherent difficulty of carrying out advanced operations in environments such as Second Life, or a lack of appropriate digital literacies. That is,

researchers or instructors may not know how to carry out various actions in VWs themselves, and as a result do not consider getting their students to engage with the VW in any meaningful way (see Molin, 2017).

Findings from Duquette and Hann indicated that a novice-expert relationship was formed between learners as intermediate level learners helped the beginner learners to complete tasks. This relates to a study by Rama et al. (2011) who noted that similar interactions occurred in an MMO environment. However, Duquette and Hann note that requiring learners to do complex operations in virtual environments seemed to result in communication breakdowns which could not be solved leading to the eventual abandonment of tasks. One implication from their study is that it is important to identify a virtual world that does not overload students' cognitive capacity and to design tasks with sufficient support so that they can be completed.

In a more rigorous investigation of the effects of multimedia on learners' cognitive load, deHaan and Kono (2010) used Sweller's (1994) Cognitive Load Theory as a framing tool for investigating how watching and playing a video game may have differing effects on learners ability to recall vocabulary from the game. Forty-six Japanese university students were placed into pairs where one student played language-based video games for ten minutes while their partner watched. Using immediate and delayed vocabulary recall tests, they found that watchers recalled significantly more vocabulary than the players suggesting that the cognitive load induced on players by the videogame was deductive to vocabulary acquisition.

Similarly, Milton, Jonsen, Hirst, and Lindenburn (2012) were concerned with assessing whether social networking sites and virtual environments are appropriate for foreign language learning and whether learning gains are afforded by participating in such communities. The study focused on university-level Hungarian learners who were paired with native English speakers in NS-NNS dyads. These pairs were asked to complete pre-determined and loosely planned conversational role-plays in several purposely-made destinations: a virtual bank, travel agency, estate agents and museum. Findings showed that compared to traditional classroom-based oral lessons, the volume of language produced in the virtual world was generally much higher. Native speakers did not dominate conversations, with learners producing on average 45% of all discourse. The opportunity for

vocabulary acquisition was also deemed high as students used the text-based chat channel to check spellings and to clarify their interlocutors' utterances. One caveat of their study, however, was that the environment appeared to be lexically impoverished and that there was a concern that without instructor intervention, there may not be ample opportunities for learners to grow their lexical knowledge to a high level of competency. This point again suggests that task design considerations are vital in determining the learning potential of such VWs.

One element that was not explored however is how the virtual world itself contributed to meaningful communication. They write that learners "spontaneously talked about many other things, such as the weather, upcoming exams and items around them in Second Life" (Milton et al. 2012, p. 108). However, there is no evidence that students were required to use specific elements of the virtual world as part of the role-play activities, their spontaneous talk appearing as tangential to the activity's primary focus. The role-plays could, therefore, have been improved. In their present form, the tasks could have easily translated to the use of Voice over Internet Protocol (VOIP) software rather than as part of communication within a VW as there was little impetus for students to interact with the environment.

Additionally, the authors claim that their study shows that VWs help generate a larger quantity of productive and genuinely communicative language by learners than is possible in a typical classroom environment, but it could be argued that the two environments are incomparable. It is not practical to have a native speaker to undertake activities with each learner in a traditional classroom environment. It seems that one of the main benefits of SCMC and particularly SCMC within a virtual classroom is access to native speakers (Chik, 2014; Levak & Son, 2017).

Finally, the authors write that manoeuvring their avatars, and interacting with the virtual world was not easy for students, and in some cases impeded on their ability to communicate effectively; one student commenting that the activities may have been easier to do if undertaken in a face-to-face environment (p. 111). The authors themselves write, "the communication in this environment is good but interaction in Second Life is still not as efficient as real-life face-to-face conversations." This relates to DuQuette and Hann (2010) and again suggests that it is essential to consider task design for VWs as there is a possible

increase in cognitive complexity due to an unfamiliarity with completing tasks in such environments.

Following, and regarding the design of tasks specifically for VWs, Gánem-Gutiérrez (2014) described pedagogical principles for the design of VW-mediated tasks. Principles were informed by sociocultural theory and in particular an activity theoretic perspective. In addition, tasks were designed to make special considerations for the utilisation of the unique affordances of the VWs. An example set of tasks is provided, which were designed in accordance with their principles. Four specific principles are focused on, illustrating how they may inform task design (based on Gánem-Gutiérrez, 2014, p. 225).

1. Learners are active participants in their learning, and therefore must be active participants of tasks.
2. Grammar emerges as a part of speech, and humans create meaning through its use. Thus, an emphasis on the need for communication between participants.
3. Producing language via languaging (Swain, 2006) and verbalising thoughts allows language to be scrutinised and leads to second language acquisition. Tasks which favour learner output and discussion must, therefore, be prioritised.
4. Instruction is critical to the development of a second language. Therefore environments which allow learners to engage in scaffolding (such as those described by the ZPD) will help drive development.

According to Gánem-Gutiérrez, these principles are to be applied in VWs that allow users to construct objects and create "places." This point then, emphasises the use of *social worlds* as a suitable domain for language learning, as opposed to the immutable *MMO* environments.

Gánem-Gutiérrez concludes her paper with tasks designed in accordance with her design principles. The example tasks (see Table 8) support learners "culturally and linguistically along a 'journey' in a virtual world" (p. 226) as they go from a presentation of L2 tense and aspect marking to pair work comparing notes to a final virtual campfire to share their experiences. The tasks are designed to be completed by NNS-NNS pairs where learners are learning each other's L1.

Table 8: Example tasks from a VW-mediated TBLT curriculum by Gánem-Gutiérrez (2014).

Week	Activity
1	Meet with your partner and watch a presentation on language forms. Discuss the areas that you understood.
2	Explore areas of Second Life connected with the L2 you are learning. Take notes and reflect on what you saw.
3	Find two or three speakers of the L2 and interview them using a set of predefined questions.
4	Prepare a picture presentation about the cultural aspects of what you discovered.
5	Practice giving the presentation to an L2 speaker and receive feedback.
6	All participants get together in the virtual world to give their presentations and give feedback. Upon completion, they may "linger around and try to get to know each other" (p. 237) making use of a virtual guitar, snacks and drinks provided.

Critically evaluating this task procedure, it is clear that Gánem-Gutiérrez has provided a robust curriculum allowing for the development of the L2 according to the four key principles. Learners are given ample opportunities to interact with native speakers of the L2 and develop their cultural understanding with these native speakers in novice-expert interactions. However, the procedure seems to be closer to a Present, Practice, Produce (PPP) approach to SLA rather than TBLT, as learners are explicitly shown a target grammatical item during the first week, and then expect learners to make use of it in subsequent weeks' activities. The task procedure also culminates in the presentation of participants' findings in a rather formal and contradictory position to the informal learning proposed by the literature review and guiding principles.

This presentation phase also seems like a wasteful use of the VW's affordances as the same activity could be completed much more easily with online presentation software. Additionally, the tasks seem to make little use of the VW's affordances for learners to create objects (a point Gánem-Gutiérrez explicitly stated as an affordance of the domain) other than the passive observation of previously created contexts. In other words, the participants appear to have little agency in manipulating the world itself other than at the very end of the procedure where they have the opportunity to engage in an informal, post-task activity: interaction with a virtual guitar and "snacks." There is room for improvement here in terms

of task design that allow learners to manipulate the environment they are situated within.

Related to the concept of task design for VWs, Hampel (2006) also provides a robust model for designing technology-mediated tasks. This model is specific to task development for synchronous online environments (such as that used in this study) and consists of three components: approach, design, and procedure (Table 9). *Approach* refers to the theories that underpin the design process, and the affordances of the technology; *design* includes the syllabus and task types as "task-as-workplan;" and finally, the *procedure* is concerned with how the task is perceived by learners as "task-in-progress" (Breen, 1987).

Table 9: Hampel's task-development model.

Approach	SLA theories
	Sociocultural principles
	Affordances of online environment
Design	Function of tasks within course
	Syllabus
	Type of tasks
	Learner/tutor roles
Procedure	Implementation in the classroom

The above studies outline how researchers are exploring the implementation of TBLT methodology in VWs. Whilst there are few studies to date which explore this area of CALL, of those that do exist, I argue that they tend to utilize open-ended, discussion tasks that do not take full advantage of the VW's immersive and interactional affordances (see also Jauregi, Kuure, Bastian, Reinhardt, & Koivisto, 2015). One way the current study could improve on Milton et al.'s findings (2012) is by designing tasks following the task design principles and procedures as outlined by Gánem-Gutiérrez (2014) and Hampel (2006). Such principles should be followed to design tasks that specifically require learners to manipulate or interact with elements of the virtual world to see if such tasks affect learner output. Doing so would also help validate the findings of Jauregi et al. (2011) also. Without the specific need for interaction with the VW environment, VOIP-based or even chat-based tasks could (and should) be used instead, as such modes place less cognitive demands on learners, and less task-creation demands on educators.

2.7 POTENTIAL BENEFITS OF CMC

Research has empirically demonstrated that chat-based interaction in CMC environments provides the same opportunities for interaction and feedback as face-to-face interactions (Lee, 2002; Darhower, 2002; Toyoda & Harrison, 2002; Iwasaki & Oliver, 2003; Satar & Özdener, 2008). Sauro (2011) also writes that chat-based SCMC has potential benefits for language learners at the syntactic, discourse, grammatical, lexical, intercultural level. Additionally, Alastuey (2010) concluded that oral SCMC is beneficial for pronunciation as it promotes the same kind of breakdown in communication as face-to-face communication, thus leading to noticing gaps in learners' interlanguage and phonetically modified output. There is thus a large volume of work on the cognitive benefits of CMC, particularly from an interactionist perspective to SLA.

Affective benefits include the lowering of anxiety in learners, particularly in synthetic immersive environments (Lee & Pass, 2014, Melchor-Couto, 2017) and an increase their willingness to communicate (Reinders & Wattana, 2014). However, the majority of this work is based on the exploration of chat-based SCMC. Chat-based CMC such as that used in chat software, instant messaging apps, and bulletin boards is considered a hybrid modality, exhibiting features of both written and spoken language (Roed, 2003; Smith 2005). The use of chat-based interaction in SCMC studies appears abundantly, and oral interaction is focused on much less (Sykes, 2014).

When oral interaction is the focus of studies, the context is predominantly in distance language education and void of the use of video (e.g., Develotte, 2009). Classroom-based studies investigating the use of SCMC as a means to develop L2 proficiency are thus the exception (Yanguas, 2010). One reason for the lack of focus on oral SCMC is that technical barriers prevent its implementation. These barriers range from bandwidth limitations, network restrictions, and screen resolution compatibility (Hampel, 2010). However, with the notion of Moore's Law (1965) being a reality (Moore estimated that the number of transistors on integrated circuit boards would double every two years, meaning that computers become increasingly faster and, as a result, cheaper), these limitations have become much less apparent.

One caveat of text-based SCMC compared to face-to-face interaction is that it lacks

pragmatic information, such as gestures and voice intonation. It is also important to emphasise that the two modes differ both structurally and stylistically, where written modes tend to be more formal and complex, and spoken modes are more informal and structurally simpler (Satar & Ozdener, 2008). Additionally, few studies explore the communication of NNS-NNS pairs as is common in monolingual classrooms around the world. Native speakers are often employed as interlocutors due to the affordance of SCMC to connect learners with L2 speakers (Milton et al. 2012). This is especially evident in studies that utilise MMOs (Rama et al. 2011). The current study hopes to fill the gap in the literature regarding the use of VWs for oral communication by assessing whether there is a benefit for low-level monolingual learners in a classroom environment.

As a critique of CMC studies, some of the potential benefits of SCMC appear to be promoting the innate affordances of the technology, while ignoring the importance of language pedagogy. One concrete example of this can be seen in Satar and Özdener (2008, p. 596). In their literature review, they cite another paper (Cheon, 2003) as providing a rationale for the use of SCMC in classroom settings:

> CMC could prove to be an efficient tool in providing more time for speaking practice, especially in crowded or teacher-oriented classrooms. Cheon (2003) reported the results of her study in such a Korean English as a foreign language (EFL) context and points to the importance of synchronous CMC (SCMC) activities, during which "individual language learners receive [sic] limited number of speaking turns." (Satar and Özdener, 2008, p. 10).

This appears to be an oversight of literature on language teaching (from a CLT perspective and beyond). What are the authors' assumptions of classroom-based language learning pedagogy? Without the use of SCMC, I argue that there are ample opportunities for learners to engage in language learning activities and authentic interaction if the instructor implements such activities. Thus, the mere introduction of SCMC does not logically increase chances for interaction if the classroom is "teacher-oriented" or the pedagogy employed by the instructor is not aligned with an interactionist perspective to SLA. In other words, if a speaking activity could have been carried out face-to-face but is not, why would the introduction of SCMC in such a context immediately open up the avenue for "providing more

time for speaking practice?" I argue for a more nuanced approach to the use of CMC in classroom contexts based on curricular goals, student needs and teachers' familiarity with and knowledge of technology.

2.8 THE EFFECT OF MODALITY ON LEARNER PERFORMANCE

As the current study explores the difference in oral performance between VW-based SCMC and face-to-face settings, this section provides details on previous studies which explore how the medium can affect learners' oral proficiency.

Warschauer (1995) compared the way second language learners communicate when using computers and face-to-face interactions. The specific aims of the paper were 1) to examine whether the quantity of learner output was more balanced during CMC compared to face-to-face interaction, 2) assess who benefits from computer-mediated communication based on gender, nationality, age and language proficiency, 3) correlate learner motivation towards participating in CMC with amount of output and 4) examine whether language used during CMC sessions was more complex than face-to-face sessions.

A pre-test questionnaire was used to assess 16 participants personal background, attitudes towards using computers, gender, age, nationality, and the number of years studying English. Students were randomly put into one of four groups and given conversation topics to discuss for 15 minutes. Two groups would discuss face-to-face, and the other two would use CMC. After the first conversation finished, groups changed their mode of communication. A post-test survey was used to understand student attitudes towards using CMC. Findings showed that students generally participated more equally when using the computer mode of discussion and that of the four nationalities present, Chinese, Japanese and Vietnamese students produced more output when communicating electronically. The Filipino students did not show any increased output. Student attitudes towards using computers to communicate were generally positive, where "their attitude toward electronic discussion was slightly better on the average than that toward face-to-face communication" (p.16). CMC also produced significantly more complex language use. Results of this study thus represent one of the earliest findings on the benefits of SCMC.

Following, Beauvois (1997) compared the oral test scores of two groups of learners that engaged in discussion sessions via CMC and FTF. The FTF control group had discussion

sessions in the classroom, and the CMC experimental group conducted the same discussions via text-based SCMC. Test scores showed that the experimental group outperformed the control group. The study, therefore, demonstrating the possibility of skill transfer from a written chat environment to oral performance. Similarly, C. Blake, (2009) found that a text-based SCMC group outperformed both FTF and blended learning conditions on oral proficiency tests after a six-week experimental period, hypothesising that text-based SCMC allows for the successful automatization of lexical and grammatical knowledge.

Finally, related to the effect of text-based SCMC on learners' oral proficiency development, Satar and Ozdener (2008) conducted a study exploring the effects of SCMC on learners' speaking proficiency and anxiety. The study compared three groups. Two experimental groups conducted several extracurricular communication tasks in pairs based on the task categorisations in Pica et al. (1993). The two experimental groups used text (chat) and voice-based SCMC tools. A control group was also employed, which did not do any extracurricular tasks. Results suggested that while both of the experimental groups showed language proficiency gains, reaffirming that text-based SCMC can help promote oral proficiency (Abrams, 2003), only the text-based group showed a decrease in anxiety levels. They thus propose that text-based CMC offers the affective benefits of improving learners' confidence by providing them with a safe environment to practise using the L2 and evaluate their performances.

Similarly, Yanguas (2010) explored learners' oral output as they undertook tasks in video SCMC and audio-only SCMC. Fifteen dyads were separated into three groups: video oral SCMC, audio-only oral SCMC, and face-to-face communication. Negotiation for meaning occurred in all three modes, yet there was a difference in the amount between the video and audio-only groups. The video SCMC group negotiated for meaning in the same way as the FTF group, but the audio-only group were forced to make use of linguistic resources to compensate for the lack of visual stimuli. One caveat, however, is that although the audio-only group produced more language, this did not appear to lead to successful negotiation.

The main finding of the paper was that task design was most influential in determining the type of negotiation focus. The tasks employed were seeded with 16 unknown

lexical items, and thus, unsurprisingly, negotiation triggered by lexical items appeared the most. Yanguas, therefore, agrees with Pica et al.'s (1993) argument that the negotiations are highly sensitive to the tasks used. Additionally, turn-taking patterns in the oral SCMC mode were shown to be equivalent to FTF patterns, but opposite to those found in written synchronous CMC.

Another study which provides contradictory evidence to the claims of the CH is Baralt (2013) who examined the effects of simple and complex tasks on participants' oral interactions in FTF and SCMC contexts. They concluded that greater task complexity led to increased L2 development in the FTF context, but simple tasks led to better results in the SCMC context. In other words, there was an incongruity between the two modes of communication focused upon; task complexity improving performance in FTF contexts but hindering performance in SCMC contexts. Similarly, Nik (2010) found that simple tasks used in a written (text-based) SCMC context led to improved accuracy but not complexity, suggesting that increased task complexity may not facilitate L2 learning in SCMC contexts in the same way that it does in FTF contexts, at least in terms of facilitating accurate output. This study therefore possibly discredits the claims of Robinson's CH which claims that increased task complexity should promote an increase in output accuracy and complexity. In other words, it may be hypothesised that the CH is only applicable to FTF contexts. The present study aims to provide further evidence for this claim.

Abrams (2003) compared the oral production of an FTF control group with two different experimental groups: written ACMC, and written SCMC. The ACMC group engaged in a week-long discussion via an online bulletin board and the SCMC group used a WebCT chat tool for 50 minutes during class time. Observations suggested that there was a higher quantity of output from the SCMC group, but no significant differences were found among the groups regarding output quality. Quality was assessed in terms of lexical richness, diversity, and syntactic complexity. These findings were also mirrored in a study by Fitze (2006). Fitze compared two types of student conferences, written and face-to-face, in terms of textual features and participation. Results showed that there was no statistically significant difference in the number of words produced for the two modes of communication, but that the quality of the written mode was higher than the face-to-face mode in terms of lexical range. Students participating in the written mode also demonstrated a higher level of

interactive competence. It may, therefore, be hypothesised that the lack of time pressure of text-based SCMC allows learners to focus more attention on the accuracy and complexity of their output, something more difficult with the time pressures of oral modes.

De Marco and Leone (2013) looked specifically at the different usage of discourse markers by learners of Italian when using computer-mediated and face-to-face modes of communication. The computer-mediated communication element of the study was undertaken using VOIP software. The participants were three intermediate-level university students paired with native Italian speakers. Conversations were generally undirected and lasted for 10 minutes in each setting. The conversation topic was different for each encounter. Findings showed a difference in the frequency of discourse markers used by the lower level learners, but no difference was seen with the more advanced learners. Specifically, the lower level learners used more discourse markers in face-to-face communication. As deHaan & Kono (2010) found, the implications of this study are that for low-level learners, the cognitive load of working via CMC could be higher, resulting in less comprehension. Issues with audio quality could also have caused the low-level learners to check what they heard more frequently.

Zalbidea's (2017) study explored the effects of task complexity and modality on learner L2 performance. The main task of the study was for participants to decide which of five hotels best met a list of requirements. Following Kuiken & Vedder (2011), task complexity was controlled by manipulating the number of elements variable. In the -Complex task, participants had three requirements, and in the +Complex task, they had six requirements to fulfil. The modes of communication compared were written ACMC (via e-mail) and spoken (monologue, to an imaginary interlocutor). The study was concerned with measuring learners' syntactic complexity, lexical complexity, accuracy and use of conjunctions.

Results of the study suggest that task complexity played less of a role in determining learner performance than modality. The speaking task promoted learners to produce more syntactically complex output, whereas the written mode promoted more lexical complexity and accuracy (p. 343). Task complexity parameters were not statistically significant in promoting output complexity or accuracy. Accuracy scores were also higher for the written

mode of communication. In conclusion, Zalbidea writes that overall, task complexity produced marginal differences in performance, but task modality played a substantial role in determining performance along all of the dimensions of L2 output focused on for the study.

While Zalbidea's study provides useful insights into the effect of task complexity and mode of communication on learner performance, the modes compared were considerably different. Due to the editability of prose with the written mode, it is perhaps logical to expect that learners' accuracy is higher than the written mode task (see also Satar & Ozdener, 2008). The tasks were also monologic, void of any interaction with an interlocutor. Therefore, while helping to form predictions regarding the effect of task complexity on learner performance, the current study does not employ a written mode, but instead two oral modes which could mitigate the substantial effect of modality found here.

As well as the effect of modality, there are several empirical studies which explore task *type* and learner performance (often with a focus on occurrences of negotiation for meaning). The following section focuses on those SCMC studies which specifically explore the effect of task type on learners' output.

2.9 THE EFFECT OF TASK TYPE ON LEARNER PERFORMANCE IN SCMC CONTEXTS

R. Blake (2000) compared the potential of two task types for producing episodes of negotiation for meaning in learner output: *jigsaw* and *information gap* tasks. The study found that jigsaw tasks promoted more negotiation for meaning than the information gap tasks, as theorised by Pica et, al. It was also found that negotiation for meaning was triggered mostly by confusions around lexical items. His conclusion was that jigsaw tasks allow for form-focused instruction, and specifically the development of vocabulary knowledge. However, due to a paucity of syntactic negotiations in the data, it was assumed that grammar development might not occur through the completion of tasks alone, thus implying the need for a post-task focus on forms phase.

Smith (2003) also explored the effect of task type on negotiation in a chat-based SCMC environment using a total of four tasks: two jigsaw tasks and two decision-making tasks. The decision-making task type was chosen as a contrast to the jigsaw task due to Pica et al.'s (1993) hypothesis that this task type should elicit less negotiation for meaning. Similar to Blake, he found that negotiation for meaning was triggered by lexical problems. However,

unlike Blake, findings revealed that decision-making tasks promoted more negotiation for meaning than the jigsaw task type. Smith concluded that this could be due to task design employed in the study. The decision-making tasks were seeded with unknown lexical items that were deemed important to understand in completing the task. For the jigsaw tasks, however, Smith posits that participants may have relegated the target lexical items lower level of importance, thus not focusing on them less.

Peterson (2006) conducted a similar study to Smith, exploring the types of interaction management that occurred when learners completed three different task types in the virtual world *Active Worlds*. The three tasks employed were jigsaw, information-gap and decision-making. Additionally, learners' interactions were chat-based. Results appeared similar to those of Smith's study in that negotiation for meaning was mostly triggered by lexical issues. Also, the decision-making task elicited the most negotiation for meaning among learners.

Finally, a study by Jee (2014) also lends weight to the notion that decision-making tasks promote more negotiation for meaning than jigsaw tasks when being conducted in VWs. She investigated how ESL students negotiated meaning in Second Life as they completed jigsaw, decision-making, and opinion-exchange tasks. This study differed to those of Smith and Peterson by analysing learners' verbal interactions, rather than the text-based alternative. Findings suggested that, like those studies that precede this one, problems with lexical items triggered negotiation for meaning, and that decision-making tasks elicited more negotiation for meaning than the other task types. One reason given for this finding is again due to task design, and the necessity of specific lexical items for task completion (Smith, 2003). Finally, and a comment regarding how little the affordances of VWs are utilised in VW-mediated TBLT studies, Jee seemed to fall short in effectively utilising the affordances of Second Life, particularly in the jigsaw task. She writes (p.10): "Students used very limited functions of the avatar movements and gestures, and no avatar movements were observed during [sic] Jigsaw task." This was due to the jigsaw task design, which only required learners to look at a handout, rather than the computer monitor.

In summary, although Pica et al. hypothesised that jigsaw tasks should promote the most negotiation for meaning and thus opportunities for SLA development, there is conflicting data in the literature on SCMC, which shows that both jigsaw tasks and decision-

making tasks may provide the most opportunities for negotiation for meaning. The current study does not focus specifically on negotiation for meaning but instead explores the effect of modality on learners' oral complexity, accuracy and fluency. Based on the studies outlined above, one task-type that has not been explored as fully is *information-gap*, a task type that seems particularly suited to support language development with low-level learners due to the type of interaction they promote – the turn-based provision of information from one interactant to the other in a systematic, linear format. Additionally, as posited by Pica, Kang, and Sauro (2006), information gap tasks promote interaction that orients learner attention to form, function, and meaning.

2.10 META-ANALYSES REGARDING THE EFFECTIVENESS OF CMC FOR LANGUAGE ACQUISITION

The above sections introduced specific, individual studies that explored task design in VW-mediated TBLT, the benefits of CMC for language acquisition and the effect of task-type on learner output. Finally, this section highlights the findings of meta-analyses concerning the effectiveness of CMC for language acquisition, with a focus on learner output.

One particularly comprehensive meta-analysis was conducted by Lin (2014) which comprised of 25 studies selected from a period of 12 years (2000 – 2012) that were concerned with the impact of CMC on students' oral proficiency. Her study synthesises all of the work that has been done in this field over the past decade, providing both a general summary of findings and revealing gaps in the literature. Additionally, her findings are a useful indicator of what results may be found with the current study. More than half of the studies included (56%) used VOIP services, while approximately one third employed text-chat (36%); and only two studies used both modes. Concerning temporality, seventeen studies (68%) employed real-time synchronous communication, three (12%) adopted delayed asynchronous communication tasks, and five (20%) adopted both modes of communication.

A wide variety of tools were used to explore the effect of CMC on oral proficiency. These range from researcher-developed platforms, free chatroom facilities provided by *Skype*, discussion forums, and class management systems such as Moodle. Looking through the papers cited reveals that no studies utilising VWs were included in the meta-analysis. According to this meta-analysis, the default tool used for verbal communication appears to

be VOIP services such as Skype.

Task types employed in the 25 studies are:

- 18 opinion exchange
- Two information gap
- One decision making
- One jigsaw
- Two mixed task types

The study which utilised a jigsaw task generated a negative effect on oral performance. However, this study was not concerned with CAF measures in particular, but rather pronunciation (Alastuey, 2010). Opinion-exchange studies also seemed to produce the smallest effect size. This finding indicates a concentration of research on the effect of CMC in the potentially least facilitative task type rather than those deemed to be most facilitative for SLA (i.e. jigsaw, information gap, and problem-solving). Reasons for this are not given. However, I argue that opinion exchange tasks were selected due to the affordances of the medium used. As mentioned above, the default tool for verbal communication was VOIP software, which can be considered a way to put two (or more) interlocutors face to face (albeit virtually), and as such, the interactive affordances available are generally considered to be conversation only. Of course, it is possible to create jigsaw or information gap activities for use with VOIP software, but this seems to have been overlooked in the majority of studies.

Generally, the results of Lin's work revealed that communication mediated by technology produced a moderately positive effect on L2 learners' oral proficiency compared to FTF communication or with no intervention. However, CMC is seen to have a detrimental effect on participants' accuracy and fluency. There is no specific result for complexity, and, according to Lin, the scoring strategy most adopted by researchers is a holistic approach, and does not make specific references to the literature on complexity accuracy or fluency measures. When individual components were assessed, fluency and accuracy appeared as the two most common. There is also no mention of studies concerned specifically with language complexity and CMC.

Subsequently, and worth mentioning here, is that reading aloud seemed to elicit the best oral performance from learners. This result comes from Lord's (2008) investigation into

the use of podcasting as a way to improve the pronunciation of Spanish learners. While Lord's study shows that podcasting seemed to be efficacious, it is far removed from the current study. Participants were not asked to converse with an interlocutor but instead, read several scripts aloud to an asynchronous audience. Interaction with other participants was relegated to a different mode: written comments on each other's podcast recordings. Lin's meta-analysis, therefore, exemplifies the breadth of research that is categorised under the term "computer-mediated communication."

Finally, Lin writes "Tasks designed for a traditional F2F environment *might* need to be modified in order to accommodate technological features" (p. 135, italics added). I argue that task modification is compulsory, and should accommodate specific features of adopted technologies. Alternatively, if educators do not modify tasks to utilise the affordances of specific technologies, they should reconsider their initial reasons for adopting them. Indeed, González-Lloret and Ortega call for such when designing technology-mediated tasks:

> "The development of pedagogic tasks should take full advantage of a chosen technology to do what cannot be done in the classroom with paper and pencil: integration of multimedia for rich, authentic input […] and engagement in learning by doing that allow students to use the language and the technology in productive and creative ways." (González-Lloret and Ortega, 2014, p.8)

In summary, the findings of Lin's meta-analysis are only somewhat relevant to the current study in that she conducted a meta-analysis over such a broad range of CMC related research.

Additionally, there are several issues with this meta-analysis. Firstly, oral production is not explicitly defined as either monologic or dialogic. Secondly, learning gains for oral production is similarly broad, including pronunciation as well as complexity, accuracy, and fluency measures. Finally, there does not appear to be any studies in this meta-analysis that are concerned with the use of virtual worlds as a domain for communication, emphasising the gap in the literature on CMC for such contexts.

Cerezo, Baralt, Suh and Leow (2013) investigated studies that use e-tutors and SCMC for L2 development. E-tutors are described as "software that allows learners to practise independently, without the help of a teacher or a peer" (p. 295), thus technology-mediated

instruction void of human interaction. Of 16 studies, seven were on the developmental effects of e-tutors, five studies compared e-tutors against FTF instruction, three investigated the effectiveness of SCMC, and one study compared the effectiveness of SCMC to FTF. They concluded, "at this point, no strong argument can be made about whether or not the medium matters in L2 development" (2013, p. 294) and that the effectiveness of CMC or FTF communication seems to depend on several possible variables such as task-type, task-complexity and modality.

The meta-analysis, therefore, goes against the findings of Zalbidea (2017) where modality was considered a significant influence on performance. What is different here is that the meta-analysis has included SCMC (as opposed to ACMC, which was the CMC modality used in Zalbidea). This hints to the idea that if modalities have the same interactional requirements of learners (e.g. face-to-face and oral SCMC both requiring learners to speak), then the effect of modality is minimised. The current study, comparing two oral modes may help shed light on this claim further. One such study which helps back up this claim is Yanguas (2010) who, as mentioned above, found no difference in the volume of negotiation for meaning episodes between video-based oral CMC and face-to-face modes. However, it is still unknown how VW-based oral SCMC compares to FTF communication. Considering both modes require oral interaction, one may speculate that communication features are similar. The current study aims to investigate this concept.

One final meta-analysis relevant to the present study is Ziegler (2016a), who conducted a meta-analysis of 14 studies that explore the relative effectiveness of SCMC and FTF communication on learner production. The study was framed from an interactionist perspective to SLA (Gass & Mackey, 2007), meaning that among several factors, studies for this meta-analysis were selected based on whether there was a comparison of interactions in SCMC and FTF environments. Additionally, only studies of a repeated measures design were accepted, i.e. studies where participants completed both SCMC *and* FTF interactions. The findings of Ziegler's meta-analysis indicate that in general, both modes of communication had a significant impact on second language development. More specifically, interactions via SCMC seemed somewhat more beneficial than FTF in supporting learners' development of written skills, whereas FTF communication seemed more beneficial than SCMC in supporting the development of oral skills. This result is unsurprising when considering that

FTF communication in the studies was only conducted orally and writing as a skill only appearing in studies which utilised SCMC.

Additionally, the advantage for performance on oral measurements in FTF contexts was minimal with an effect size of only $d = .04$. One reason considered for the advantage of the FTF context in supporting the development of oral skills is gesture, a feature only available in FTF and video SCMC contexts. Again, this relates to Yanguas (2010) who found that audio-only SCMC produced more episodes of negotiation for meaning over the video-supported SCMC, where he hypothesised that the lack of visual information of one's interlocutor (gesture, facial expression, etc.) caused more breakdowns in communication.

2.11 Summary of CMC studies

CMC has been researched from multiple perspectives due to 1) the broad range of available modalities. There are, however, two main modes: synchronous or asynchronous, plus several modalities: oral or chat-based, with and without access to video, and finally as part of activities within digital games or virtual environments. Research on CMC is also conducted to discover its effect on various linguistic functions such as pronunciation (Alastuey, 2010) and negotiation for meaning (Smith, 2003; Jee, 2014) and affective variables such as anxiety (Baralt & Gurzynski-Weiss, 2011). Specific considerations for task design have also been made by several researchers (Gánem-Gutiérrez, 2014; Hampel, 2006). Findings suggest that task type plays an influential role in determining output, where decision-making tasks have been shown to elicit more negotiation for meaning than jigsaw tasks (Jee, 2014; Smith, 2003). Affectively, CMC may help reduce foreign language anxiety due to the amount of time available to create utterances (Satar & Ozdener, 2008) or through the use of an avatar (Reinders & Wattana, 2014), yet anonymity provided by avatar-usage may not be the only factor that reduces anxiety (Melchor-Couto, 2017).

In keeping with the current study, oral SCMC has been shown to help produce gains in oral proficiency (Satar and Ozdener, 2008), and improve pronunciation in communicative contexts (Hung & Higgins, 2016). In studies that employ a VW as the domain for oral SCMC, benefits range from promoting interpersonal interactions and greater output than in classroom-based work (Lan, 2014) and providing authentic experiences of the foreign language (Chen, 2016). However, studies that compare the performance of learners as they

complete tasks in VWs versus FTF are missing from the literature. Kim (2017) explored the effects of task modality on learner performance and recognised this dearth, calling for further studies to compare learner performance variability over oral FTF, oral SCMC, and text-based SCMC modes (p. 14-15).

2.12 CHAPTER SUMMARY AND RATIONALE FOR THE CURRENT STUDY

This chapter initially outlined the author's stance towards instructed SLA, highlighting the claims of the Interaction Hypothesis and its connection with tasks and TBLT. Definitions of a task were provided, and the psycholinguistic underpinnings of TBLT were introduced including a section on how tasks may affect learners' oral performance in terms of complexity, accuracy and fluency. The Cognition Hypothesis and the Limited Attention Capacity Model were presented, the two prevalent models which help predict how task conditions such as cognitive complexity may affect learner performance. Based on findings from studies exploring task type and output, task design considerations were also made.

Next, the technology employed in this study was introduced via a historical account of its development from plain text-based worlds that were stored on large networked computers at university laboratories to immersive 3D environments available on individuals' personal computers. Virtual worlds and their more ludic cousins "MMORPGs" were also introduced before outlining how such digital technologies may be categorised as part of DGBLL. This included the three main research approaches: game-based, game-enhanced and game-informed language teaching (Reinhardt & Sykes, 2014). This overview helped position the current study as concerned with *game-based* language learning. Studies related to both game-enhanced and game-based language learning were then introduced demonstrating the L2 learning potential of virtual worlds and other digital domains from both cognitive and socially informed perspectives.

Following that, studies which explore CMC and learner performance were reviewed in detail, where several gaps were identified. Few studies focus on learners' spoken performance during VW-mediated tasks, studies comparing learners' output over multiple modes of communication have focused primarily on text-based interactions, and VWs rarely feature in CMC-related studies. Finally, the benefits of virtual worlds for language learning in monolingual contexts has also received little attention to date, justifying the need for the

current research.

In answer to the foundational question of whether there is value in conducting research on VWs and L2 communication, the following additional rationale was considered: With the rapid sophistication and availability of communication technology, research on CMC has seen a progression from asynchronous to synchronous, and, in a somewhat more partial way, from text-based to spoken modality. As an extension to the continuum of CMC technology, VWs represent the next technological "frontier" and a gap in our knowledge regarding how communication in these domains may differ to both face-to-face and other CMC modes. Although our collective use of virtual worlds is far from being as normalized as mobile phone use for communication, it is not difficult to imagine a future in which virtual world-based or, indeed, virtual reality (VR)-based communication becomes a norm (however, whether such a future is a utopia or dystopia is not an argument made in this dissertation). Therefore, as technology becomes increasingly embedded in the lives of language learners, the demand for suitable curricular and pedagogical implementation of such technologies is considered paramount. This dissertation explores one such avenue of inquiry: the integration of VWs for instructed SLA, along the dimension of what Gonzalez-Lloret and Ortega consider "technology-mediated" TBLT (2014).

2.13 RESEARCH QUESTIONS

Based on the aims outlined in the introduction, and the gaps in the literature exposed in the literature review, I now present the research questions for this study.

1. What is the effect of modality on learners' oral task performance as they complete tasks of increasing complexity via two different modes of communication?
2. What are the affective affordances of completing tasks in a VW compared to the FTF equivalents?

These questions are formulated to help fill the current gap in the literature regarding learners' spoken output while undertaking tasks in virtual worlds.

Research Question 1 is concerned with the effect of the interaction between modality and task complexity on learners' oral task performance in a task-based language teaching context. It could also be reformulated with a focus on task complexity. That is: "What is the

effect of task complexity on learners' oral task performance?" However, as the two variables are intricately linked, forming the two factors for a two-way repeated measures analysis, both modality and task complexity will be analysed simultaneously. The research question is explored via the use of three sets of tasks. Task conditions for the three sets of tasks are manipulated in order to create tasks of differing inherent task complexity. It is, therefore, possible to investigate whether there is an interaction between modality and task complexity on learner output.

There is a dearth of research exploring how VW-based SCMC may affect learners' oral output, particularly in comparison to an FTF mode. Framed from Robinson's Cognition Hypothesis and with reference to studies that explored the effect of modality on learner performance, a hypothesis may be generated that the VW tasks will pose a greater level of cognitive complexity than the FTF counterparts, therefore either hindering learner performance (Baralt, 2010; Böhlke, 2003), or pushing them to greater oral complexity. The alternative is that there is no difference in learners' oral task performance based on modality or task complexity manipulations. In this case, differences in performance could be attributed to learner-related factors or pedagogical interventions. Such has been found in Sauro (2012) and Baralt (2013) and hypothesised by Skehan (2016).

Research question 2 is concerned with uncovering the effect of modality and task complexity on learners' motivation towards studying English. While there are already several studies which explore the affective benefits of learning with digital games or virtual environments (e.g. Baltra, 1990; deHaan, 2005; Reinders & Wattana, 2015), they are often not compared to another modality. In this study, I explore the affective affordances of the VW in comparison to the more traditional FTF mode of instruction. Additionally, the effect of task complexity coupled with modality on learners' motivation is explored.

The contributions that this study makes to the existing research on technology-mediated TBLT may be summarised as the following:

1. *Comparison of three task pairs manipulated to be of differing task complexities.*

Unlike 85% of studies on cognitive task complexity which only compare a single task manipulated along one task condition (Sasayama, 2015), this study adopted a 2 x 3 experimental design using three different task pairs. This allows for a more thorough

investigation of the effect of *modality* and *task complexity* on learner task performance as they complete multiple tasks of varying complexity.

2. *Oral task performance measured in terms of complexity, accuracy and fluency.*

Whilst some studies have focused on only one dimension of task performance such as the volume of negotiation for meaning (R. Blake, 2000; Iwasaki & Oliver, 2003; Yanguas, 2010), provision of feedback (Baralt, 2013), fluency (Abrams, 2003), or complexity (Lintunen & Makila, 2014), the current study aims to understand how modality and task complexity manipulations affect task performance in terms of complexity, accuracy, and fluency (for more examples, see Liao & Fu, 2014; Ziegler, 2016b), thus allowing for a more complete comparison of the claims of the CH and LACM regarding task complexity and learner oral task performance.

3. *Questionnaires utilised to validate task complexity predictions and provide further insight into learner task preferences.*

Studies which investigate the use of VWs as domains for language learning are often conducted to discover the affective affordances of the medium (Melchor-Couto, 2017; Reinders & Wattana, 2011, 2014, 2015). In this study, while the effect of modality on participants' motivation is explored, the questionnaires main purpose is to gather information in understanding how tasks' cognitive demands affected participants' perceptions of task difficulty, thus focusing on the validation of the assumed relationship between task complexity and task difficulty perceptions. This is achieved via the analysis of both quantitative and qualitative data, allowing for data triangulation and further validation of any findings.

3 METHODOLOGY

3.1 RESEARCH DESIGN

This study is an evaluation of the use of virtual worlds in a task-based language teaching context. Framed from an interactionist approach to SLA, and with reference to Robinson's Cognition Hypothesis, it aims to discover if there are any differences in learners' oral task performance as they undertake tasks in both virtual world and face to face modalities. This is achieved via the measurement of 20 low-level EFL learners' output complexity, accuracy and fluency. A 2 x 3 factorial design was used to investigate the influence of two independent variables on participants' oral task performance: task-complexity, and modality. All participants completed six tasks in total. A summary of the tasks can be seen in Table 10.

Table 10: A summary of the tasks used in this study

Virtual world	Face to face	Task type
Room creating	Room drawing	One-way information gap
Direction giving	Direction giving	One-way information gap
Spot-the-difference	Spot-the-difference	Two-way jigsaw

This study is framed from a mixed methods approach to research. Bogdan and Biklen (1992) argued that quantitative and qualitative methods of inquiry can both be used together and that doing so is often desirable. According to Creswell (2014), mixed methods research originates from the 1959 study by Campbell and Fisk who used multiple methods of inquiry to test the validity of psychological traits. Their approach inspired others to start mixing methods, and soon, qualitative methods of inquiry such as interviews started appearing alongside quantitative methods. Subsequently, the triangulation of quantitative and qualitative collected data appeared (Jick 1979). Triangulation is thus considered a way of compensating for the limitations and biases that occur when using any single method alone. In this way, the data collected from the two approaches may be used to reinforce findings. Indeed, Bogdan and Biklen (1992) argued that quantitative and qualitative methods of inquiry can both be used together and that doing so is often desirable.

Mixing methods is not without its critics, however. Some researchers argue that each

of the two approaches requires different procedures and have different epistemological implications. Thus, by combining both research methods, we are mostly ignoring the epistemological and ontological assumptions, or separate worldviews, associated with each method. This argument forces researchers to reconsider whether epistemology and method are inseparable or not. Bryman (2015, p. 636) writes that although Kuhn (1970) argued that paradigms are incommensurable, it is not clear that qualitative and quantitative research are paradigms. A similar debate appears in the field of linguistics between scholars who argue for a cognitive or social-cultural theoretic view of SLA (see Hulstijn, Young & Ortega, 2014)

Johnson and Onquegbuzie (2004) highlight the dispute that wages between positivists and interpretivists, who argue that their research paradigm is ideal, and thus advocate implicitly for the *incompatibility thesis* (Howe, 1988), thus positing that qualitative and quantitative research methods are incompatible. Following, Johnson and Onquegbuzie introduce and explain why mixed methods research is a natural complement to the two traditional research approaches. Their main argument is that both research methods are important and useful and that by combining them, we may achieve something more significant than we could when only utilising a single method. Finally, they provide a detailed overview of mixed method research design including an eight-step model. 1) determine the research question, 2) determine if a mixed model is appropriate 3) select the research design 4) collect data, 5) analyse the data, 6) interpret the data, 7) legitimate the data, 8) draw any conclusions. Additionally, there are eight different approaches in their typology of mixed methods research depending on three specific variables: Objectives, data collection and data analysis.

With reference to the literature review (Chapter 2), it appears that all studies which explore learner performance based on the CAF model employ a positivist approach and thus the collection and analysis of quantitative data. Accordingly, the current study will also collect and analyse quantitative data in order to accurately align findings with the broader body of literature. Various quantitative measures were chosen to explore the relationship between modality and learner output in terms of complexity, accuracy and fluency. Post-task questionnaires were utilised to collect quantitative data in the form of Likert-like scales. This data was gathered in order to gauge learner perceptions of task difficulty and cognitive load, which was then used to explore the relationship between task complexity and learners' task.

However, in order to further investigate and validate explanations of any findings gained from the quantitative data, an interpretivist perspective was also adopted. Transcriptions of learners' task performance were referred to and analysed qualitatively. Additionally, learners' responses to open-ended questions on the post-task questionnaires were analysed both quantitatively (via a custom coding scheme) and qualitatively in order to further validate any findings.

3.2 PARTICIPANTS

This study was conducted using two instances of an intact, elective class at the researcher's teaching context. Both classes were assigned as experimental groups, and there were no control groups used. Thus, convenience sampling procedures were applied in drawing a sample for this study. Their ages ranged from 18 – 21, and all students were non-native speakers. They are a homogenous group of native Japanese learners. Although they have had English education since they attended junior high school, thus receiving an approximate eight years of formal English education, they may be described as low-level learners. They receive three hours of English education each week in other classes. Reference to their in-house English proficiency test revealed that their English ability was generally in the range of a TOEIC score between 300 and 400.

3.3 INSTRUMENTS

3.3.1 Needs analysis

Long (2005, 2014) places great importance on conducting a needs analysis (NA) as part of a TBLT driven curriculum in order to discover not only what types of task students need, but the linguistic and culturally specific elements they will NA is seen as a useful precursor for materials development.

The current context can be considered an example of "teaching English for no obvious reason" (TENOR) (West, 1994). I am part of a small humanities department in a larger science and technology university. The controlled classroom experiment introduced here was conducted as part of a class which serves the humanities teachers in allowing them to explore research and teaching related to their professional interests. Students are therefore exposed to subjects that they would not otherwise encounter as part of their major route of

study. The students that enrol for this particular class know beforehand of its experimental design, the tools involved, and the types of activities they will be required to do. They enrol after reading the syllabus, which is provided to them on two ways: 1) posted on the universities internal learning management system and 2) given to students at the start of the year in a paper handbook. Based on informal observations, participants join the class based on three key factors: an interest in English, an interest in video games, or because they require additional credits in order to proceed to the following year of their studies. In this way, there is generally not a common linguistic level or motivational characteristics shared by participants other than their language learning experiences to date (as mentioned in section 3.2 on participants). In such contexts, a needs analysis is often avoided due to the perceived lack of any common needs (González-Lloret, 2015a).

Additionally, as outlined above, in the case of the current study, students are made aware of what technology is being used in the class and its experimental nature. As a result, tasks were chosen not based on identified future needs of the learners, but with reference to Robinson's Cognition Hypothesis and Pica et al.'s taxonomy of tasks (1993), providing learners with a number of tasks that vary in complexity and thus investigate Research Question 2: whether task-type, and, more specifically task complexity, is influential on participants' oral performance. Additionally, task selection and design were conducted with reference to relevant studies on TBLT and virtual environments. The tasks selected as part of this study are introduced in detail below, but first, I turn my attention to the technical skills required of participants in this study.

3.3.1.1 Technological skill requirements

Linguistic elements and task type make up part of a needs analysis, but equally important is to conduct a "skills audit" (Bax 2011) to determine the digital literacies of participants (Shetzer & Warschauer, 2000), and what technical skills they need in order to complete technology-mediated tasks (González-Lloret, 2014). Digital literacies include (1) *computer literacy*: a knowledge of basic operations of the hardware and software, (2) *information literacy* such as the ability to gather and manipulate information from the internet, (3) *multimedia literacy*, which refers to learners' ability to manipulate multimedia (sound, video, audio, etc.), and (4) *computer-mediated literacy*, which refers to technology-mediated communication and participating in online discourse practices. An audit of the technological

skills learners need in order to participate in this class was conducted where the following proficiencies were found and catered for via specific on-boarding activities at the start of the course. A summary of activities is provided in Table 11.

Table 11: A list of technical skills required as part of participating in this study.

Lesson component	**Technical skill requirements**
Pre-task	Accessing the internet Use of search engines Understanding of PC hardware in order to use headphones Logging into and navigating an LMS Understand how to click links to view content.
Task	Extensive use of a virtual environment including: • Logging into the VW • Controlling an avatar • Navigating the environment • Manipulating the environment Extensive use of computer hardware and software including: • Mouse and keyboard use • Headset connection • VOIP software usage • Use of "push-to-talk" system for oral communication
Post-task	Logging into and navigating an LMS

Both a physical and web-based, interactive worksheet was provided at the start of each lesson which introduces learning goals, sample conversations, task completion guidelines, and any follow-up language instruction details. This was accomplished with the obsolete "Chalkup" learning management system (LMS) which is now part of Microsoft "Teams" education package. Thus, learners were required to be able to interact with and navigate the LMS. Subsequently, pre-task listening activities often required students to use the internet to access online audio recordings or videos of native speakers carrying out similar tasks. The main task element of the class required the use of a specific virtual world, and

post-task activities were completed online within the LMS ecosystem.

Several lessons were dedicated to fostering the required skills necessary for this study. Participants were given detailed instructions on how to log into the virtual world, and how to use *TeamSpeak*[3] for oral communication. Knowledge of the virtual world (item names, avatar manipulation, and general controls) was fostered in a separate instructor-led class which was framed as an exploration into the virtual world. These three classes thus made up a scaffolded introduction to the tools used in the study. Due to the large amount of virtual world content, it was impossible to provide students with an exhaustive guide of the environment. However, participants were given homework to explore the game world and write a journal of their findings. This extracurricular element acted as a way to foster their knowledge and familiarity with the environment further, before participating in tasks designed for this study.

3.3.2 Choosing an appropriate VW

Based on the literature review, it was considered important to reduce any cognitive load caused by the technology used. Doing so may help promote a focus of attention on language rather than technical aspects. Additionally, as I am not technically proficient in computer programming, any VW employed in the study needed to be sufficiently simple enough for me to be able to create interactive tasks for participants. In order to best select an appropriate VW, then, the following considerations were made:

- Content should be easy to create
- Controls must be simple
- The graphic user interface should be uncluttered and intuitive
- Specialized vocabulary should be limited
- The environment should be closed to the public (to prevent outside distractions and influences on the participants' task performance)

Although MMOs feature heavily in the literature on VWs and language learning, upon conducting a review of those available, it became clear early on that they were not suitable domains for the following four reasons.

[3] https://www.teamspeak.com/en/

3.3.2.1 High cognitive demands

Beginner-level EFL learners may find that the cognitive demands of playing in an MMORPG's complex environment hinder their L2 production (see deHaan, 2005b). This is due to a number of factors, most prominently the following two: 1) the steep learning curve required to understand how to play the game, and 2) the specialised discourse used in such games. Speaking of WoW in particular, there is a highly polished and comprehensive in-game tutorial to guide new players through the basics (Peterson, 2016). However, regardless of how detailed and helpful a tutorial might be, the enormity of the game means that players will still have a lot to learn even upon completion of this tutorial. On-boarding learners to use WoW was therefore considered to require too much time; time that could be better spent on language learning tasks.

Regarding the specialised discourse used by players in MMOs, as reported by Steinkuehler (2006), even native English speakers who lack prior knowledge of MMO gaming norms may have trouble deciphering it. An example of typical in-game discourse (from the game Lineage) is presented in Table 12.

Table 12: An example of specialised discourse in MMOs (Steinkuehler, 2006 p.42).

The original utterance, translation and gloss	
Original Utterance	afk g2g too ef ot regen no poms
Literal translation	away from keys got to go to Elven Forest to regenerate no mana potions
Gloss	Just a minute. I have to go to the Elven Forest to regenerate. I'm out of mana potions.

Attending to inaccuracies in spelling, abbreviated expressions, and specialised words, as well as attempting to play the game may present too large a cognitive demand for beginner learners. A study by Peterson (2011) found that novice Japanese learners—similar to the context here—experienced difficulties trying to manage the communication systems in an MMORPG. This factor, coupled with problems in utilising game-specific commands, prevented the above learner group from engaging in beneficial types of interaction.

3.3.2.2 Complex game text

Thorne et al. (2012) analysed the quality and complexity of the language used in

game-generated "quest texts" with several measures. Their results showed that such quest texts contain "a high degree of lexical sophistication, lexical diversity, and syntactic complexity" (p. 290). Specifically, with the use of a D-Level test—designed to measure syntactic and structural complexity, and not merely proxy measures such as sentence length or word length—"the most complex level of sentences (d-7) [occurred] with greatest frequency" (p. 291). In summary, then, they conclude, "as far as the quest texts are concerned, WoW presents to players an environment that includes a substantial volume of highly complex linguistic input" (p. 291). For participants in the current study, such highly complex quest text may be considered a hindrance and not a source of useful input.

3.3.2.3 Limited task types

Activity theory (AT) is the framework used in research on the sociocultural theory which considers both physical activities, and high-level motivated thinking, doing and being of an individual in a social context (Ryle, 1999). Although the current study is not specifically framed from a sociocultural perspective, a reference to AT is useful in explaining how different virtual environments have different, specific affordances for language learning. Drawing a parallel between TBLT and AT, tasks may be considered 'artefacts' that mediate language learning through interaction. Additionally, a distinction may be made between 'task' and 'activity,' task referring to the workplan that is given to learners (i.e. the artefact), and activity as the communication during task performance. Learners, however, interpret the workplan differently, and thus the same task can result in very different kinds of activity (see Breen 1989).

From an AT perspective, then VWs may be described as a material mediating tool for learners to realise their goals (linguistic or non-linguistic). As an analogy, just as a hammer is used to drive a nail into a wall to hang up a picture, virtual worlds (and the tasks designed within them) can be used as a tool to help learners achieve the goal of language acquisition. Related to this notion, Nardi (2010) writes of WoW: "its potent agency [is] a particular kind of medium that engages players in *certain kinds of regulated performances*" (p.62) [italics added] and "the culture of a VW is enacted through human conversation and designed objects that mediate activity" (p.18). The implication being that the specific design of VWs (the mediating tool) influences the activities that players will perform, and thus, from a language learning perspective, may influence and limit the type of language that learners are exposed

to.

On this premise, the affordances of MMOs and social worlds may be assumed to be inherently different due to differences in design philosophy. From a language learning perspective then, how could these differences in design manifest themselves in the linguistic output of users? In the case of MMOs, and again, focusing on *WoW,* there are thousands of quests, but analysis reveals only a few common quest types (WowWiki, 2016). The linguistic requirements of users may, therefore, be considered limited. As a specific example, one type of task in WoW (or *quests* as they are known) requires users to kill a certain number of enemies, collect a specific item that those enemies drop and then return a certain, predefined number of those items to a non-player character (NPC).

If a few, pre-determined task types (such as that described above) make up the bulk of an MMOs' overall in-game tasks, and the literature reports that MMOs are a successful and important domain for language learning in the 21st century (see Peterson, 2016), I argue that the potential for language learning may be much higher in social worlds where teachers have the opportunity to design tasks and content, thus the opportunity to create a much larger repertoire of language-learning specific tasks. This is one of the main reasons for rejecting MMOs in favour of other VWs for the current study. One specific example of a teacher-defined task that is creatable in a social world but not generally available in MMOs is the classic information gap activity where important information required for task completion is divided between interlocutors engaged in a task. Such "gap-like" considerations are not found in MMOs, where all players have access to the same information from the start.

3.3.2.4 Content creation and content appropriateness of "social worlds."

Turning away from MMOs and focusing on social worlds now, I argue that there are several hurdles that prevent educators from being able to commit to employing these environments in their studies. Focusing on Second Life in particular, these include its complex graphical user interface, which takes time to learn and thus an extraneous cognitive demand placed on learners, and the steep learning curve for both playing and creating content (Wiecha et al., 2010).

Another issue with online social worlds is how appropriate the content is for learners. Although Second Life features a rich assortment of diverse activities for players to engage

in, trends show that players generally gravitate towards two major pastimes: shopping and sex (Nardi, 2010). This is certainly not the case for the entire game world; however, in an informal conversation with DuQuette on 30th July 2013, he confirmed that it is not uncommon to be confronted with explicit nudity and adult language in public areas. In a study by Hansen, Berente, Pike, and Bateman (2008, p. 74) involving the analysis of written responses of senior business executives who spend time playing Second Life, one response claimed, "In my 40 years of life I don't think I have ever run into so many sexually-motivated characters."

Other observations included numerous reports of harassment, stalking, and even sexual violation. Such content is inappropriate for a classroom or educational setting. Additionally, unless a researcher purchases a private area within the Second Life virtual world, there is a chance that participants, may be interrupted mid-task by players from outside the study and thus contaminating or making collected data unusable.

3.3.2.5 Reasons for choosing "Minecraft."

Having rejected MMOs as suitable domains for this project, and researching the potential benefits of several virtual worlds including Second Life and OpenSimulator, the more recent and simpler domain "*Minecraft"* was finally considered the most suitable as it matches all of the above criteria:

1. Content and language-learning activities can be easily created in a building-block fashion.
2. The graphics, user-interface, and gameplay are all simple, with almost no learning curve involved in order to achieve basic competence.
3. Servers are privately hosted; meaning that content is curated and monitored easily. Additionally, in contrast to Second Life, running the server is virtually free, and teachers have full moderation over content and users.
4. There are a plethora of player-developed *plugins*, which may be added to the standard multiplayer version of Minecraft, which allow for a great deal of world customisation allowing for the creation of a suitable environment for educational purposes.

The current study, then, makes use of a COTS game, yet I have positioned this study as *game-based* (as opposed to *game-enhanced* or *game-informed*). This requires explanation.

The three terms were introduced in Section 2.5.3.1, but as a review, *game-enhanced* refers to the use of COTS games, appropriated for language learning purposes, while *game-based* refers to the specific development of games for promoting language acquisition (based on Reinhardt & Sykes, 2014). The reason this study is positioned as a *game-based* approach is due to the unique affordances of Minecraft.

The selected "game," Minecraft, features a number of different ways in which it can be played. One such option available is a *sandbox* environment (i.e. something similar to the definition of a *social world* like Second Life). A sandbox game is a term used as an analogy to a child's sandpit/sandbox where they are free to mould the sand into any shape they want. Thus, in a sandbox game, players are free to play as they please. These games often shun narrative elements and linear gameplay (campaign) modes in favour of world exploration and personalisation instead. As such, it is incorrect to label Minecraft merely as a COTS game, but more accurately as a COTS game within which lies the option for players to appropriate it as a *social world* environment (as defined by Warburton, 2009). In more detail, then, the tasks created for this study do not exist in the original version of the game but were specifically designed by me.

Furthermore, none of the "game-like" elements found in the original version of the game were used as part of this study. This was done in order to position the study firmly in the *game-based* field. One caveat, however, which would exist whether the researcher used Minecraft, OpenSimulator or any other user-customizable VW, is that learners were required to use some environment-specific lexical items such as "crafting table," "furnace," and "slab."

3.3.3 Task design

Tasks were selected based on their affordances for promoting oral interaction between interlocutors, where considerations were made based on Pica et, al. (1993). As such, information-gap or jigsaw tasks were employed as these tasks are theorised to promote the most interaction between interlocutors. Two-way jigsaw tasks are considered the most beneficial in promoting interaction between interlocutors; however, as the participants in this study are low proficiency level, it was considered advantageous to divide the roles of speaker and listener in order to reduce cognitive load for a number of the tasks employed. Although some studies of SCMC have found that decision-making tasks promoted greater negotiation

for meaning (Smith, 2003; Jee, 2014), researching how tasks may promote different levels of negotiation for meaning is not the target of the current study, and thus the former task types are employed. Additionally, as outlined in the pilot study section (Section 3.4.1), due to the low proficiency of the participants, tasks with closed-goals were chosen in order to limit the amount variability in task performance such as task completion time.

Three pairs of tasks were created for this study. Based on task complexity operationalisations and by referencing Robinson's Triadic Componential Framework for task conditions, the pairs of tasks were categorised as low, medium and high complexity. There are also two modalities employed in the study, virtual world SCMC and face-to-face communication. Abbreviations for the tasks are used throughout the rest of the dissertation. For example, the low complexity face-to-face task is named FTF LOW, and the high complexity VW task is VW HIGH. Following is a detailed description of the tasks used in this study.

3.3.3.1 Environment creation

One major hurdle to conducting this research was creating the interactive VW tasks. This section outlines the technical aspects and proficiencies required to create the tasks, as a way to both documents the process, and to allow for the replication of the study.

The initial step in this project was for the purchase of a physical server to host the Minecraft world. This step is not strictly necessary, as the Minecraft world can be run on a desktop or laptop PC. Regardless, once a suitable technical infrastructure is set up, educators are required to download and run the Minecraft server files. This will automatically host an instance of Minecraft on the user's local machine. There are several sources where these files may be downloaded for free (such as https://mcversions.net/). Once the server files are ran on a local PC, the instructor needs to locate and identify the IP address of the PC which other users (students/participants) must point their client. The current study required the researcher to purchase a total of 20 individual Minecraft clients (one per user). The clients were purchased at Mojang's homepage[4].

The initial stage of creating these tasks was selected in-game items and a place to host

[4] https://minecraft.net/

the different tasks. Once an area was selected, it was excavated, and buildings which I have outlined in the above sections were created; namely: the two houses for the spot-the-difference task, and the rooms for the room decoration task. Fortunately, the traversable city which was used in the directions task was not created specifically for this project but sourced and imported from a Minecraft world creation community online who offer their map for free[5].

Minecraft is highly customizable with thousands of community created plugins. They can be downloaded at various sites, the main one being the Spigot homepage (https://www.spigotmc.org/resources/). These plugins expand the functionality of Minecraft in various ways allowing for actions that are not present in the original game. Plugins were utilised heavily in the creation of tasks for this study. For an introduction to additional plugins suitable for language educators, see York (2014). Below is a list of the main plugins and how they were used to create additional functionality in this study.

EssentialsX[6] allows for warp points to be created within the world. Players can then input a command to warp to these user-defined locations. This plugin was utilised throughout the study to assemble learners at the same location quickly. For instance, after completing a directions activity, learners could input the command */warp ship* to quickly return to the start point (avoiding the unnecessary "empty" time of retracing their way back). **EssentialsX Group Manager**[7] allows for permissions to be applied to users based on the groups to which they belong. For instance, in this study the following two groups were created:

- Teacher – able to edit all blocks on the server
- Student – only able to edit blocks in specific areas.

WorldEdit [8] is a plugin which allows players with the appropriate level of permissions to take a snapshot of the world state. This snapshot is named (for instance: "pre-task snapshot") and can be recalled after students have completed tasks so that the terrain is

[5] http://oldshoes.ca/broville
[6] https://dev.bukkit.org/projects/essentials/files
[7] http://wiki.ess3.net/wiki/Group_Manager
[8] https://dev.bukkit.org/projects/worldedit

reverted to the snapshot condition. This was utilised in the room decoration task so that once the participants had completed the task, the teacher could revert all of their edits to the pre-task condition. Additionally, WorldEdit allows the server owner to set up regions which can be editable based on group permissions. For example, the area in which the two houses for the spot-the-difference task were created was demarcated as a region which required teacher-level permissions to edit. This meant that there would be no ill effect if the student-group players tried to destroy blocks in the area.

Finally, of note is the **Citizens**[9] plugin which allows the admin of the server to generate non-player characters (NPCs). This was utilized in the spot-the-difference tasks to generate the eight non-player characters. These characters had to be named and positioned appropriately around the map.

Having introduced a detailed description of how the VW tasks were created, including technical considerations, the following section moves onto considerations for measuring learners' oral task performance. The section is subdivided into complexity, accuracy and fluency measures where the rationale for their adoption is provided.

3.3.3.2 LOW task pair: Spot-the-difference

A spot-the-difference task is a closed, information gap or jigsaw task which is typically used in studies concerned with both TBLT (McDonough, 2005) and SCMC (e.g. Pellettieri, 2000; Satar & Özdener, 2008). Spot-the-difference tasks often require learners to use prepositions as they describe the layout of a room or the position of items in space. In the current study, however, there is a focus on using the present continuous form of verbs, as participants have to relay information to each other about the actions of several different characters. Compared to the directions task (see below), spatial-awareness is not specifically tested here, and thus, this task pair set may be perceived as less complex than the other two, particularly the FTF version.

Following Robinson's Cognition Hypothesis, the primary design feature of the spot-the-difference tasks was the number of elements (i.e., [+/− few elements]), operationalised according to the number of characters. There were eight characters to find in total, two of

[9] https://www.spigotmc.org/resources/citizens.13811/

which are doing the same activity in each version, making a total of six differences. This task pair is considered the lowest complexity of all three tasks pairs and is referred to as LOW throughout the results, discussion and conclusion section.

The VW version of this task features two identical houses with a divide that separates them into two spaces. The divide is much larger than that of Figure 8 and stretches around the front of the houses meaning that participants are unable to see each other's house. This prevents the learners from being able to see each other and their avatars, thus increasing cognitive demands along the (+/-) *here and* now task condition.

The characters, distributed around these traversable houses, are undertaking various, deliberately vague activities. The vagueness adds to the difficulty of the activity, as participants have to figure out what each character is doing. For the FTF version, a single screenshot with all characters undertaking similarly vague actions was used (Figure 9). Additionally, in order to increase the similarity between the two tasks in this pair, the illustrations used in the FTF version of the task features the same game-world context, characters, and items as the VW version. However, the actions characters are undertaking are different.

Figure 5: VW spot-the-difference task screenshot illustrating each participant's house.

Figure 6: FTF spot-the-difference task reference picture

For both versions of the task, participants were required to fill in a worksheet as they discover whether their characters are doing the same activity or not (see Table 20). Switching from the game world to the real world to complete the worksheet may thus be considered an additional cognitive demand of the VW task.

Table 13: An example of the spot-the-difference task worksheet to be completed by participants.

Character Name	Your picture	Your partner's picture	Same or different?
Beeroman			
Boss			
cheapsh0t			

3.3.3.3 MID task pair: Directions

For the directions task, players use a 3D town in the virtual world (see Figure 13), and an overhead 2D map of the same town for the offline, face-to-face version (see Figure 14). According to Robinson (2001a), the complexity of such a task can affect learner fluency,

where requiring learners to both think of a destination and the route can be detrimental to fluency. In the current study, all destinations are predetermined to reduce such cognitive demand. Additionally, participants take it, in turn, to direct their interlocutor, thus making this another one-way information gap task repeated for each participant.

Figure 7: An example of the 3D town for the online task.

Task complexity was operationalised by requiring participants to give a fixed number of directions each. For example, participants would guide each other to areas of the map that were close together, or opposite a location, their partner had directed them previously. Also, this task features an increase in complexity about the (+/-) *independency of steps* resource-dispersing task condition. Compared to the previous LOW complexity task where participants were free to discover differences between partners in any order, this task requires learners to give directions sequentially, following a set pattern. The directions task pair is therefore considered to be between the spot-the-difference and room decoration task (next section) in terms of complexity. It will be referred to as MID complexity going forwards.

The VW version requires learners to move their avatar through the 3D space, which may impact fluency as they wait for their partner to reach specific waypoints or destinations. Both VW and FTF versions of this task provide learners with planning time to locate the start location and the locations to which they may have to guide their partner. The maps feature waypoints (the white numbers in Figure 14) which may be referenced during task

performance. The VW task provides learners with the opportunity to navigate the virtual terrain and locate their individual before starting the task. In this way, the paper map is not referenced during task time, keeping a focus on the immersive 3D environment.

Figure 8: An example of the overhead 2D map for the FTF task.

3.3.3.4 HIGH task pair: Room Decoration

Referencing DuQuette and Hann (2010), an information gap activity that requires learners to rearrange objects in accordance with predefined models was employed. This task draws participant attention to grammatical forms, comparative and superlative adjectives, naming objects, question raising, describing, and prepositions. The closed-goal nature and specific aims of the task should help promote fluency, and complexity may be controlled based on how many elements participants are required to interact.

In order to create an information gap, the VW specific version of this task used a 3D space separated into two areas, one for each participant. These areas were further divided into two rooms. Each room was pre-built to allow one participant to give instructions to their partner (see Figure 8 for an example).

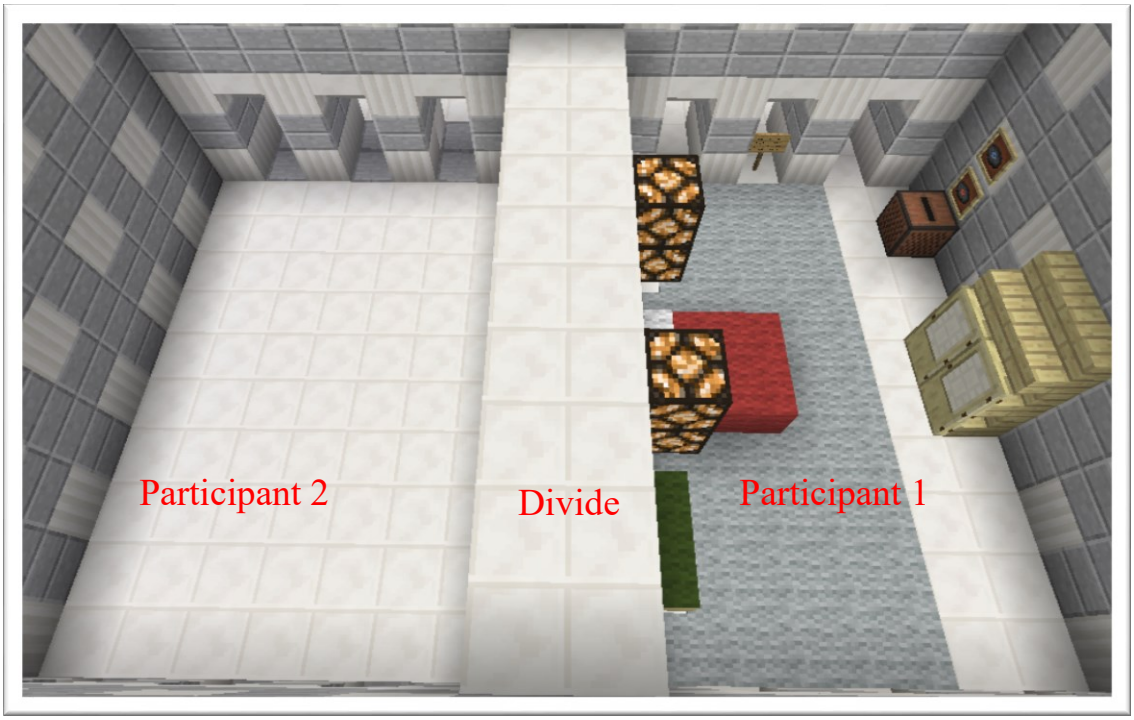


Figure 9: An example of the online decoration activity.

Participants take it in turn to experience the instructing (speaking) and building (listening) roles (Table 14). Thus, this task may be classified as a one-way information gap task repeated for each participant. For the face-to-face version, the same task was conducted with pictures on paper, where learners had to draw the objects themselves (Figure 10 and Figure 11). Again, in order to keep the two tasks as similar as possible, the offline FTF version of the task also featured items from the virtual world.

Table 14: Room layout for VW information gap activity.

	Player 1	**Player 2**
Room 1	Empty	Completed kitchen
Room 2	Completed bedroom	Empty

Again, following Robinson's Cognition Hypothesis, there were two primary design features of this task pair. Firstly was the number of elements, operationalised according to the number of objects in each room. As can be seen in the example figure below (Figure 10), there were nine objects which participants were required to describe to their partner for each room (Two furnaces, a sink, a tap, water in the sink, carpet on the furnaces, a window, a cake,

and a crafting table). The second element was the level of (+/-) *spatial reasoning* that this task required of learners. Unlike the spot-the-difference tasks, the room decoration tasks require learners to create something based on their interlocutor's instructions within a 2D or 3D environment (FTF: draw in 2D, VW: place blocks in 3D). One consideration for this task-pair, then, was to keep the number of objects to be manipulated the same for each mode of communication. Finally, this task may be considered to place an additional cognitive demand on learners in the form of the (+/-) *perspective-taking* resource-directing task complexity condition of the CH as learners have to give directions to their partner with the consideration of how they perceive the scene, thus having to take on their perspective (if it assumed that perspective-taking here refers to the physical rather than metaphorical). How this differs from (+/-) *spatial reasoning* is however unclear. The naming convention used for the six task complexity resource-directing dimensions in Robinson's CH have been criticised in that it is unclear how they may be operationalised (D. Ellis, 2011).

Based on the above researcher-determined operationalisations of task complexity along the resource-directing dimension, and the inherent cognitive demands of this task type, the room decoration task is considered to be the most complex of all three task pairs, therefore this task is referred to as HIGH throughout the results, discussion and conclusion section of the dissertation.

Figure 10: Participant 1 offline room decoration speaking example.

Figure 11: Participant 2 offline room decoration listening example

3.3.3.5 Task complexity predictions

In this section, I refer to the LACM and CH in order to predict task complexity for all six tasks used in this study. I initially consider task complexity at the task pair level (LOW, MID, and, HIGH) and then the tasks within each task pair, thus how modality may affect task complexity perceptions. A summary is provided listing tasks in order of most to least complex.

3.3.3.5.1 Task complexity predictions for task pairs

Based on the CH, this section formalises task complexity predictions for the three task pairs.

The room decoration task pair (HIGH) is considered the most complex of the three due to having the most number of elements to manipulate (each participant is in charge of both manipulating objects and describing objects) the most spatial reasoning required, and the most steps needed to complete the task. This is followed by the directions task (MID), and finally, the spot-the-difference task (LOW), which is considered the least complex. See Table 15 for a detailed overview of task complexity considerations for each task pair. Predictions made here will be validated in two ways: 1) indirectly by analysing participants' oral task performance and comparing this with Robinson and Skehan's hypotheses, and 2) directly, by asking for participants' perceptions of task complexity with post-task questionnaires, gathering subjective notions of *task difficulty*.

Table 15: Complexity comparison of the three task types utilised in this study.

	Spot-the-difference	**Directions**	**Room building**
Cognitive/conceptual demands (resource-directing)			
Number of elements	Low	Medium	High
Spatial reasoning	Low (Characters are only in a single place)	Medium (there are multiple ways to get to the same location)	High (items have to be placed precisely according to a set model)
Perspective-taking	Low	High	High
Performative/procedural demands (resource-dispersing)			
Planning time	Medium	Medium	Medium
Few steps	Low	Medium	High
The predicted complexity level of task pair			
	Low	Medium	High

Task difficulty differs from *complexity*, as defined by Robinson (2001a), and is a subjective variable based on students' level of proficiency, experience with the L2 and in the case of this study, technical expertise. Assessment of task difficulty is achieved via participant feedback. In order to gather data on participants perceived difficulty for each task then, a Cognitive Load Subjective Experience Questionnaire was employed (as seen in Paas, 1992; Paas, van Merriënboer, & Adam, 1994; deHaan & Kuwada, 2010).

In terms of task conditions, and first of all concentrating on *participation* conditions, the tasks were all designed to be completed by dyads, thus keeping the variable *participant number* equal. The only variable that provokes further discussion is the number of contributions needed to complete each task, which appears to match the assumptions of task complexity above: the spot-the-difference tasks was operationalised to require the fewest contributions and the room decoration tasks the highest. There is also a great deal of homogeneity between learners in this context, which mitigates the majority of *participant* conditions. Participants are of generally similar proficiency, were predominantly male, of

similar age, and equal social status.

3.3.3.5.2 Task complexity estimations per modality

The above section provided a hypothesis regarding the complexity of task pairs, where spot-the-difference tasks were operationalised to be of low complexity, the directions tasks as medium complexity and the room decoration tasks as high complexity based on manipulations of (+/-) number of elements and (+/-) spatial reasoning. However, what of tasks within a task pair? Does modality influence task complexity according to the criteria laid out by the Cognition Hypothesis? I argue that it does in the following ways.

Although VWs are considered immersive environments where communication resembles real-world (FTF) communication more closely than other SCMC modalities, in terms of the (+/-) here and now condition, VW tasks still require learners to be separated from each other within the classroom context, indicating that the (+/-) here-and-now condition is lessened for the VW tasks, and therefore a higher cognitive demand.

Additionally, the spatial reasoning of traversing a 3D virtual landscape is more pronounced than the FTF mode. Consider the room decorating task for example. With the FTF version, participants' view of the scene is fixed. There is no need to consider their interlocutors current perspective and orientation to the room. More concretely then, the concept of direction, be it "left" or "right" is the same for both participants. However, and in contrast to this, in the VW a participant's frame of reference is not fixed, and again, as an example, "left" could be interpreted as forwards, right or backwards based on participants' current orientation to the scene. This is an additional complexity for the VW tasks.

3.3.3.5.3 Summary of task complexity predictions

Having considered task complexity at both the task-pair and modality levels, Table 16 presents the prediction of task complexity for each of the six tasks used in this study. In terms of task pairs, the room decoration tasks are predicted to be the most cognitively demanding, followed by the directions tasks and finally the spot-the-difference tasks. For tasks within a task pair, the VW tasks are considered to have higher cognitive demands than the FTF equivalents.

One question that remains, however, is how task complexity compares between individual tasks. For instance, although task pairs may be of different complexity to each

other (HIGH > MID > LOW), and tasks *within* those pairs differ based on modality (VW > FTF), will task complexity be perceived as a linear scale as portrayed in Table 16, or will there be overlap between the tasks? For example, it is possible that the Directions VW task is perceived as more complex than the Room decoration FTF task, or, the Spot-the-difference VW task could be perceived as more complex than the Directions FTF task. In order to negate these concerns, the data were analysed using two-way repeated measures ANOVA statistical tests (more details in section 3.5.3 below), where modality and task complexity (at the task pair level) are considered the two factors for all assessing all numerical data.

Table 16: Task complexity predictions.

Task complexity	Prediction
6 (Highest)	Room decoration VW
5	Room decoration FTF
4	Directions VW
3	Directions FTF
2	Spot-the-difference VW
1 (Lowest)	Spot-the-difference FTF

3.3.4 Measures of linguistic performance

Findings from Juaregi et al. (2011) seemed to suggest that tasks which did not require direct interaction or manipulation of the virtual environment promoted learners to speak more during task performance. However, they did not examine this finding in any formal manner, such as by counting the number of utterances for each learner, or with the aid of statistical tests on the spoken data. Additionally, a detailed literature review of SCMC studies revealed that few studies have explored task creation for virtual domains, especially in comparison to traditional face to face communication

The current study thus aims to generate knowledge in this area by analysing learner spoken output in terms of the CAF model. This method of investigation that is often employed in task design studies to measure and evaluate learners' oral (and written) task performance (examples include Robinson, 2011; Yuan & Ellis, 2003; Skehan & Foster, 1997, 1999, 2001). The following section introduces the measures used for each dimension of oral performance in this study, with rationales for their inclusion provided.

3.3.4.1 Complexity

The current study explores oral complexity in two ways: structural complexity and lexical complexity. In TBLT research the former tends to be prioritised (Geng & Ferguson, 2013). However, referencing recent literature on learner output complexity reveals that lexical complexity is now considered as important as structural complexity, and the two terms have been separated, where the superordinate category of "Complexity" now refers mainly to *structural* complexity, and *lexical* complexity exists standalone as "Lexis" giving a new abbreviation of "CALF" for complexity, accuracy, lexis and fluency (see Housen, Kuiken, & Vedder, 2012; Skehan, 2009).

In this study, structural complexity was measured by the *number of syllables per utterance*. This is similar to the *words per turn* measure outlined in Jackson and Suethanapornkul (2013). However, syllables were used as the base unit in this study because participant utterances were generally short. Measuring in terms of syllables thus allowed for a finer measurement of complexity than is possible at the word level.

Several additional measures were considered but rejected after completing the pilot study. These were: words per AS-Unit (Lambert & Engler, 2007; Revesz, Ekiert & Torgensen, 2014) and clauses per AS-Unit (Foster and Tavakoli, 2009; Hsu, 2017). During data analysis for the pilot study, it was found that the number of instances where participants used multiple clauses per AS-Unit were extremely low. This is attributed to two factors: 1) the oral nature of the discourse and 2) the low proficiency of the learners. Utterances were found to generally only contain one clause. Statistical analyses of these measures produced insubstantial results as the data was wildly non-normally distributed and statistical significances were absent. This was attributed to the lack of any significant differences in participants' performances for these measures.

Subsequently, lexical complexity was measured via two measures. The first is by quantifying the *total number of different words* used based on types. This includes verb tense, modality and voice (see Yuan & Ellis, 2003). The website *Lextutor*[10] was used to analyse data for this measure. The second measure is the Measure of Textual Lexical Diversity

[10] http://www.lextutor.ca/vp/eng/

(MTLD) test, which requires further, detailed explanation.

A type-token ratio (TTR) of learner discourse was considered, but later discarded; again, due to an unforeseen bias in the pilot study. The lowest level participants were found to use very few function words, which artificially inflated the TTR score. Results tend to show that those participants with greater English proficiency (and thus fluency) had lower TTR scores due to the higher frequency of function word usage in their samples. This phenomenon is a recognised issue in the literature on vocabulary use, where shorter texts result in higher TTR values.

A number of improvements have been proposed including the Guiraud index (Guiraud, 1960), D score (Malvern et al., 2004), and the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010). Samples collected in this study ranged significantly in the number of tokens they contained. The lowest being 60 and the highest at 1448 tokens. Although a number of the results were less than Koizumi's recommendation of a lower limit of 100 tokens, the MTLD test was considered the most appropriate second measure for lexical density. The MTLD test was conducted via the website Text Inspector[11]. The score of the MTLD shows the "mean length of word strings that maintain a criterion level of lexical variation" (McCarthy & Jarvis, 2010 p. 381). In terms of assessing the values, Jarvis and Daller (2013) provide the following example. Considering two MTLD values of .41 and .78, the .78 value indicates that the text data uses a high variety of words such that the TTR does not drop below .72 until 78 words are used. In contrast, the .41 result means that the texts TTR value drops below .72 after just 41 words are used. The .78 value, therefore, shows a greater lexical diversity than the .41 value.

In summary:

- Structural complexity is measured by
    - The average number of syllables per utterance
- Lexical complexity is measured by

[11] http://textinspector.com/

- The total number of different words used
- An MTLD test

3.3.4.2 Accuracy

In this study, accuracy was measured as the number of error-free AS-Units. This is similar to studies by Sample and Michel (2014) and Amiryousefi (2016) who used T-Units instead of AS-Units due to the output being written as opposed to spoken. However, taking cue from Yuan and Ellis (2003), morphology, syntax and lexical mistakes were also included.

AS-units were favoured over T-Units and C-Units due to the nature of the discourse. All of the tasks are oral, interactive tasks, and during the pilot study utterances were found to be highly fragmental and filled with pauses, as is often the case with oral communication. For dialogic oral data, Norris and Ortega (2009) prescribe the use of AS-units as the most appropriate measure of subordination as it contains many non-syntactic segments. The AS-unit is defined as "a single speaker's utterance consisting of an independent clause or sub clausal unit, together with any subordinate clauses associated with either" (Foster et al., 2000, p. 365).

- *She is wait for you* contains a **morphological error**
- *She waiting is for you* contains a **syntactic error**
- **Lexical errors** may refer to missing words, or the use of a word incorrectly such as
  - This is sushi shop (word missing)
  - And turn left on the first corner (misuse of 'on')

Concrete examples of each error as they appear in the data can be seen in the following three examples.

Table 17: Verb conjugation error (taken from Spot FTF task).

Utterance number	Participant Number	Utterance
93	16	My Seetricks is *ride* on the boat.*

Table 18: Subject-verb agreement error (taken from Spot FTF task)

Utterance number	Participant Number	Utterance
04	01	He *have* a record.*

Table 19: Pluralisation error (taken from Room VW task).

Utterance number	Participant Number	Utterance
04	01	But two *block* up.*

The quantity of each error type was recorded in order to determine if modality influenced the error type. Additionally, since one-word utterances are easy to produce and inflate the accuracy results, they were not included in the results. Finally, rephrased portions of speech were not marked as erroneous but excluded.

3.3.4.3 Fluency

As reviewed above, there are two predominant ways of measuring output fluency, a property of linguistic output which is concerned with the speed at which a learner can access or deploy their knowledge of the L2. Measures belong to either *temporal* or *vocal* fluency categories. In this study, fluency was assessed from a temporal standpoint as the "number of syllables uttered per minute," or, in other words, the rate of speaking (Yuan & Ellis, 2003; Geng & Ferguson, 2013). In more detail, the measure was calculated as the total number of syllables divided by the number of seconds required to complete the task and then multiplied by 60. False starts, hesitations and interjections such as "um" or "er" were included if they had appropriate meaning to the communication context. Unlike studies by Sasayama (2015) and Adams & Nik, (2014) the total number of words produced was not used a quantity-based measurement due to the large variance in task completion times between the six tasks and therefore variance in the number of words spoken.

Rate of speech was considered a valuable measure of learners' fluency as it links to an interactionist perspective to SLA. In more detail, if a learners speech rate (temporal fluency) is high because a task was able to promote them to produce a high volume of words

per minute, this means that 1) they have more opportunities to test hypotheses of the L2, 2) breakdowns in communication and resultant negotiation for meaning episodes are more likely to occur, and 3) there is more opportunity to receive feedback on erroneous utterances. Thus, a high speech rate may be considered beneficial to second language acquisition from an interactionist perspective due to its direct links to the output hypothesis.

3.3.5 Post-task questionnaires

Two questionnaires were employed in this study; an *individual* post-task questionnaire (Appendix 5 – 7) which learners completed after completing each task, and a *post-task comparison* questionnaire which they completed after finishing both tasks in a task pair (Appendix 8). Both questionnaires were given to students in their native language. Questions had been translated from English with the help of a native Japanese colleague to ensure no miscommunication occurred. The individual post-task questionnaire is designed to assess learner perceptions of task difficulty, the mental effort required to complete each task, and how enjoyable the tasks were to complete. The questionnaire was used to ascertain whether predictions regarding task complexity converge with participants' perceptions of task difficulty.

There have been numerous studies that ask participants to give their perceptions of task difficulty (e.g., Gilabert, 2006, 2007; Kim, 2009; Révész, 2011; Robinson, 2007). However, there are only a few task complexity studies that employ a questionnaire designed to measure cognitive load (as seen in the work of Paas et, al., 1994). As an exception, Sasayama's (2013, 2015) studies utilised both a mental effort and task difficulty scale, the results of which suggested that participants' perceptions of mental effort and task difficulty for four different tasks corroborated with task complexity predictions. Révész, Michel and Gilabert (2016) also propose that participant self-rating of cognitive load is necessary in order to successfully validate task complexity.

Based on deHaan and Kuwada (2010), participants were given a Cognitive Load Subjective Experience Questionnaire after each task (based on Pass et al., 1994). Four questions were used to assess their mental effort and the perceived difficulty of tasks (Appendix 5 - 7). Mental effort and task difficulty are inter-related but not the same measure because a student may perceive a task as difficult but not willing to put in the mental effort

to complete it. Additionally, as participants' attention was divided between interpreting the task goals (via paper and pencil or computer controls) and their spoken output, one item in the post-task questionnaire asked learners to reflect on where their attention was most directed, towards motor skills (drawing/controlling their avatar) and producing English sentences. This provided another way to check the cognitive load of the VW on participants' working memory.

3.3.5.1 Individual post-task questionnaire

The individual post-task questionnaires (FTF and VW versions of the same questionnaire exist) contained two sections. The cognitive load section was designed to collect quantitative data via Likert-scales. An example can be seen in Table 20 with data collected from the pilot study. The data is used to validate task complexity predictions, and to answer RQ2. That is, the fifth measure on the questionnaire asks learners to rate their enjoyment of the tasks, compared with the participants' perceptions of task difficulty, data generated from this measure will help reveal how both modality and task complexity affected enjoyment.

Table 20: An example of mean scores for the cognitive load questions.

Task	Mental effort	Task difficulty	Vocabulary difficulty	Focus – Using English or doing the activity	Enjoyment
VW LOW	3.23	3	2.38	3.62	4.15
VW MID	3.18	3.09	2.64	3.55	3.73
VW HIGH	3.89	2.95	2.79	3.53	4.32

The second part of the questionnaire featured open-ended questions. These were:

- What did you learn in today's activity?
- What were the positive points of this activity?
- What do you think should be improved with this activity? If you have any suggestions, please write them below.

The first question was designed to assess learners' perceptions of what learning gains the individual tasks promoted. Responses to this question, therefore, provided information regarding which aspects of language the participants concentrated on during task performance, and if there are any differences between the modes of communication, and, furthermore, task type. The second question was vaguely written in order to uncover both the elements of the task that learners enjoyed, as well as a self-reflection on their task performances. Finally, the third question was designed to get learners to evaluate the task design elements, thus providing insights into the particular affordances or difficulties posed by modality.

Data from the open-ended questions was referenced in two ways. First, a coding scheme was created to generate quantitative data which is then compared to the responses to the closed questions to look for correlations. Then, the open-ended responses were referenced qualitatively allowing for a more detailed understanding of phenomenon related to the research questions. That is, responses are referenced in an attempt to explain *why* certain tasks were perceived to be difficult, enjoyable, mentally demanding, etc. Eliciting open-ended responses in this way is considered an invaluable step in order to answer such questions (Sasayama, 2015).

3.3.5.2 Task comparison questionnaire

Upon completion of a task-pair (that is, both the VW and FTF version of a task-pair), participants were required to explicitly compare their experiences during the FTF and VW tasks, where they were directed to give a preference for either the VW or FTF task. Four measures were utilised on a 5-point Likert scale where 1 was assigned to showing a preference for the VW task, and 5 to the FTF task. The four questions were:

1. Which task promoted you to speak more English?
2. Which task did you enjoy more?
3. Which task was more difficult?
4. Which task do you think had the most learning potential?

The first question was designed to measure whether learners perceived modality to

be influential in promoting spoken output. The results of which are used to validate results collected from the oral fluency and complexity measures. Subsequently, asking learners to specifically state which of the two modes was more enjoyable will aid in answering RQ2, and more specifically: whether there are positive affective affordances of studying in the VW. The third question relates to the individual post-task questionnaire measure on task difficulty, however, here it is posited as a direct comparison of the two tasks in a task pair, again, to uncover the effect of modality on learners perception of task difficulty.

The final question relates to the affective affordances of the modality. Answers to this question may help map perceptions of task enjoyment to perceptions of tasks' learning potential based on modality. One of the main reasons for utilising games in educational contexts is due to their potential positive influence on learner motivation by leveraging their intrinsic enjoyment of gaming (Barab, Thomas, Dodge, Carteaux & Tuzun, 2005). As found by Allen, Crossley, Snow, and McNamara, (2014, p. 139) game enjoyment and game "helpfulness" (what the current study is referring to as "learning potential") correlated highly. Learners' perceptions of helpfulness was considered a significant predictor of task enjoyment. This alone does not guarantee that learners in their study learnt more when completing game-based activities. However, in terms of increased engagement, and therefore the potential for learning to occur, game enjoyment was considered a highly influential factor. The current study aims to replicate this avenue of inquiry by looking for a correlation between learners' perceptions of task enjoyment and learning potential with this post-task comparison questionnaire.

Additionally, space for learners to write their opinions about the tasks in a task pair, as well as reasons for their answers were provided. Similar to the post-task questionnaire, open-ended question answers were coded to look for patterns in responses. This data was referenced to validate further findings generated from the quantitative data. Additionally, answers were analysed qualitatively to explore further the reasons for any results found. This questionnaire was given to students in Japanese and was translated with the help of a native Japanese colleague.

3.4 PROCEDURE

This study took place in a science and technology university in Japan, using two intact

elective classes. The study can, therefore, be considered a controlled classroom, quasi-experimental design. Student participants were enrolled on the elective course, promoted as an experimental class on learning English with virtual environments. The class was 15 weeks long, and students met once a week for a 1.5-hour lesson. Activities used in the class were developed over multiple instances of the course, one instance of which acted as the pilot study. Students receive double credits for completing this class and are assessed based on their attendance, general attitude and level of participation, and on the submission of homework throughout the course. Their performance during tasks is not a part of the assessment criteria.

I, the researcher of this study, also acted as the instructor of the class. Such a situation can raise ethical issues such as power-harassment and forced volunteering. In order to avoid these issues, the assessment of the class was purposely separated from the requirements placed on participants of this study. In other words, the questionnaires and recorded data collected as part of this study were not used in any way to evaluate them. Additionally, participants were given a consent form at the start of the course which allowed them to decline from participating in the study.

In traditional TBLT contexts, during the task phase of the lesson, the instructor is often reduced to the passive role of time-keeper or making sure that students are on task (Willis, 1996). The same is true of the current context but to a greater extent. It was impractical for the instructor to monitor all participants at once, so instructor interference during task time was avoided. This was to ensure that the spoken data are not adversely influenced by the instructor. In the case that a participant directly asked for help from the instructor, such help was given. However, the majority of cases where I was called over to help was to alleviate technical issues rather than language-related episodes.

Participants were randomly divided into pairs to work on a total of six tasks designed to promote oral communication, three using a virtual world environment, three face-to-face. The pairs stayed intact over the length of the study. For the face-to-face tasks, students sat next to each other in adjacent seats facing each other, and for the VW tasks, they sat opposite each other with computers between them. Headsets and microphones were used for communication via dedicated software. The classroom context for the study is presented in

Figure 12.

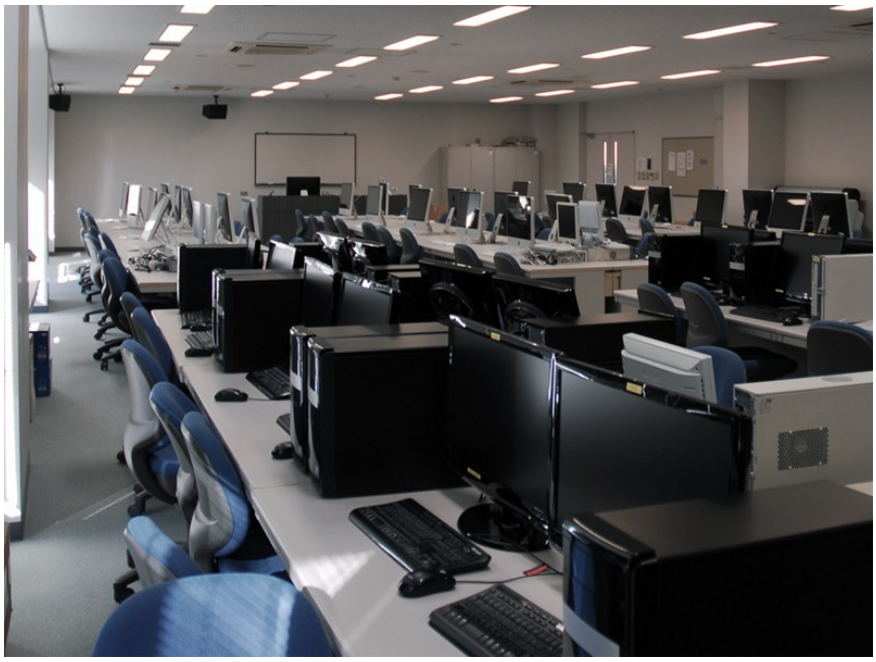


Figure 12: Classroom context for the study.

3.4.1 Pilot Study

A pilot study was conducted between April and July 2015. The aim was to familiarise myself with:

- Task creation and selection
- Data collection practices
- Transcription techniques
- Data analysis
- Test the validity of instruments

Firstly, the pilot study utilised six task-pairs (double the number of those finally selected), with a mix of both open and closed tasks. In this study, while keeping Skehan's (1998) definition of a task in mind, open tasks were defined as tasks where there are definite, obtainable goals. However, the answers needed to fulfil those goals are not predetermined. Closed tasks were defined as tasks where the teacher has created a predetermined set of "correct" answers that the students are required to complete during the task.

Many of the tasks in the pilot study were also longer in duration than those finally

employed, taking between 50 minutes to 1 hour to complete. Practical problems this created include: 1) tasks could not be completed within class time; 2) transcribing recorded data was arduous; 3) task performance varied a great deal between participants as they completed open and closed-goal tasks. As a result, certain tasks were removed from the study, and those that were eventually employed were shortened. The number of learner utterances and length of time to complete these closed-goal tasks was more consistent than open-goal tasks. When completing open-goal tasks, it was found that participants spoke much less frequently, and task completion time varied considerably between dyads. In addition, with closed-goal tasks, it was easier to create comparable tasks for the two modes of communication by limiting the requirements for task completion (such as requiring each participant to direct their partner to a specific destination twice each). These two issues led to only adopting closed-goal tasks in the final version of the study.

It is worth exploring why closed-goal tasks promoted more spoken interaction than open-goal tasks between participants in a dyad. One possible reason is that a level of L2 proficiency higher than that of the participants of this study is required to perform well at open-goal tasks. The openness of tasks has been shown to enable or inhibit production based on learner proficiency in several papers (Rankin, 1995; Lambert and Engler, 2007). Lambert and Engler (2007) compared the complexity, accuracy and fluency of students' oral output over several open and closed-goal tasks. Their results indicated that higher proficiency students benefited from the use of open-goal tasks, but they write that "closed tasks may be most useful at the novice level and of more limited use with advanced learners" (p.42).

Subsequently, there was a problem with the post-task questionnaire designed for the pilot study. Alternatively, at least, in the way that it was distributed to participants. Initially, participants were directed to complete an online questionnaire via an LMS. However, some participants did not complete the questionnaires due to technical issues or lack of computer literacy skills. In order to prevent this, it was decided that participants complete a paper-based questionnaire during classes.

It was also at this stage that the initial transcription techniques were tested. As it was difficult to find examples of other studies transcription techniques I decided to create my own using spreadsheet software to record dyads' utterances. In order to make learner utterances

as clear as possible, recordings were edited in the audio editor Audacity[12]. In particular, the noise was removed, volume compressed, and silent periods of audio truncated to reduce the amount of time needed to conduct the transcription process. Coding techniques were also considered, along with appropriate CAF measures. These are introduced in the data analysis section below (section 3.5.1).

3.4.2 Lesson procedure

A set structure for all lessons was adopted. This was to ensure that instruction was the same each week, keeping the variance in instruction to a minimum. For example, planning time was kept approximately the same, a variable which can influence output complexity or accuracy (Mehrang & Rahimpour, 2010; Skehan, 2016). The teacher's role during task time should also be mentioned. As students undertake a given task, the researcher monitored their conversations but did not provide corrections.

Each class is comprised of the following stages:

- Pre-task
    - Introduction to topic
    - Brainstorm useful vocabulary based on the topic to activate their current knowledge.
    - Watch or listen to native English speakers carry out a similar task and take notes.
- Task
    - Participants carry out the lesson's task in pairs.
- Post-task
    - Teacher-led focus-on-form session.
    - Participants completed cloze-gap questions.

[12] https://www.audacityteam.org/

Regarding the focus on forms section, language forms that appear during the task phase are pre-selected based on the pilot study. These forms are focused through several activities including cloze versions of the native speakers' text, specific instruction on grammatical forms, and questions designed to focus student attention to such forms. The focus on forms activity is conducted partly in class time and partly as a homework assignment. An example worksheet used in class can be seen in Appendix 4.

Additionally, as Sample and Michel (2014) discovered, with growing task-familiarity students can focus their attention on all three CAF dimensions simultaneously, which has direct relevance to the current study in that the language goals are the same for both VW and FTF tasks in a task-pair. Although modality is different for the two tasks, they could conceivably be interpreted as two instances of the same task. Thus, to ensure that familiarity does not affect results, the order that participants completed tasks was different for each dyad, in a counterbalanced configuration. An overview of the order that tasks were completed can be seen in Table 21. The modality of the first task in a task pair was alternated, and the order of completion was reversed for each class.

Table 21: Class number and task completion order.

Task	Class 1	Class 2
Spot-the-difference	FTF → VW	VW → FTF
Directions	VW → FTF	FTF → VW
Room decorating	FTF → VW	VW → FTF

In summary, then, Table 22 provides a full overview of the procedure followed in this study:

Table 22: Complete overview of both classes.

Week	Class 1	Both Classes	Class 2
1		Pre-study questionnaire to gauge technology proficiency and gaming history	
2		Technology orientation 1: Introduction to the project, VW, LMS and communication software.	
		Homework to learn more about VW.	
3		Technology orientation 2: Navigating the VW together as a whole group.	
4	Spot-the-difference FTF		Spot-the-difference VW
		Post-task questionnaire	
5	Spot-the-difference VW		Spot-the-difference FTF
		Post-task questionnaire	
		Post-task pair questionnaire	
6	Directions VW		Directions FTF
		Post-task questionnaire	
7	Directions FTF		Directions VW
		Post-task questionnaire	
		Post-task pair questionnaire	
8	Room decorating FTF		Room decorating VW
		Post-task questionnaire	
9	Room decorating VW		Room decorating FTF
		Post-task questionnaire	
		Post-task pair questionnaire	

3.4.3 Data collection

Participants' speech data was collected via audio recordings. When undertaking VW tasks, the dedicated software *TeamSpeak* was used to save a recording of interactions to the participants' PCs. Each dyad was assigned a number and was asked to name the audio recording file on their PCs with their dyad number included. For the FTF tasks, an audio recorder was utilised. The recorder is placed between both participants as they undertake tasks. Each recorder is numbered, and participant dyads had a recorder assigned to them in order to match recordings with participants. Recordings from both sources were transferred to my secure PC straight after each class.

Questionnaires were administered post-task. As mentioned in the above section on the pilot study, these questionnaires were initially given to students in electronic format via an LMS. However, due to inadequate completion rates in the pilot study, for the current study participants were given paper-based questionnaires at the end of each class to complete during class time. This consideration was considered particularly important in increasing the validity of their responses, as participants were able to complete the questionnaires having just finished the task, meaning their experiences could be called upon easily and referenced accurately.

3.5 DATA ANALYSIS

The following section outlines the totality of data analysis techniques employed in this study from transcription and questionnaire coding considerations to the rationale behind employing various statistical analysis techniques.

3.5.1 Transcription considerations

An example of transcribed data is shown in Table 23. The coding applied to each utterance was created by the author. Transcription spreadsheets were divided into two sections. Utterances are divided based on pauses, or the interactional nature of the discourse. To the right of the utterances are a number of codes. These codes are described below.

Table 23: An example of transcribed data

Utterance Number	Participant Number	Utterance	Fluency & complexity code (Syllables)	Accuracy codes
55	07	Ah, OK.	2	
56	08	Where do I need to go	6	y
57	07	OK.	2	
58	08	I want to go to Thomas's house	9	y
59	07	Ah, OK.	2	
60	07	OK	2	
61	07	At first go up	4	y
62	07	Go over the bridge	5	y
63	07	And turn right.	3	y
64	08	Mhm.		
65	07	And go up to the stair.	6	l m
66	07	OK.	2	
67	07	At the top of the stair you can see...	9	m
68	07	You turn right.	3	l
69	07	Turn right and go straight the road.	7	l
70	08	OK.	2	

Accuracy codes employed were: Y(es) for an accurate utterance, L to indicate a lexical error, M to indicate a morphological error and S to indicate a syntactic error. If more than one type of error occurred in the same utterance, then multiple code letters were written in the same cell. For example, the following utterance contains all three error-types and would be coded as L M S:

- You can see right big buildings*
- (You can see a big building on the right)

Fluency was indicated as a syllable count next to the utterance. Certain rules were followed when creating syllable counts. In many cases, learners would repeat the same word multiple times. Typically, this was for confirming what their interlocutor said with the general expression, "OK." In the situation where the student said "OK" multiple times, only the first instance was counted. The reason for this is that the two alternatives change both the fluency score and the *syllables per utterance* measure used for complexity. The two alternatives that were rejected are, 1) count all instances of "OK" as a single utterance ("OK, OK, OK" equalling a syllable count of six) which artificially inflates the *syllables per minute* fluency score, or 2) count each instance of "OK" as a separate utterance, which then reduces the *syllables per utterance* complexity score. Thus, counting only one of the repeated words was considered the best compromise for keeping fluency and complexity scores accurate. This is a common practice in fluency studies as seen in Ellis (2009), and Lintunen, & Mäkilä (2014). The additional rule that was employed to determine the number of syllables per utterance was only to count those syllables for English words.

For the complexity measure *syllables per utterance*, a participants' average syllables per utterance score was recorded on the sheet by dividing the total number of syllables (collected from the fluency column) with the utterance count for each participant. The MTLD test was achieved by exporting all utterances for a particular learner to the website "Text Inspector" and analysed there. Similarly, the website Lex Tutor was utilized to generate data for the different words measure. The results of these tests were placed on additional sheets of the spreadsheet for each dyad's performance.

As a review, for each participant, the following quantitative data was collected per task.

- Complexity measure
    - Syllables per utterance
    - Different number of words (using Lex Tutor)
    - MTLD (using Text Inspector)
- Fluency measure
    - Syllables per minute
- Accuracy measures
    - Accurate utterances (%)

- Lexical errors (%)
- Morphological errors (%)
- Syntactic errors (%)

3.5.2 Coding responses to the post-task questionnaires

Participant comments were analysed qualitatively for emerging themes and categories. These were then clustered into superordinate categories based on common themes. For the individual post-task questionnaires these superordinate categories were comments that related to 1) the task, 2) language or learning, and 3) the environment. Each of these categories contained a number of subcategories as seen in Table 24 below.

Table 24: Coding scheme employed for comments to the open-ended questions of the individual post-task questionnaire.

Task	**Language (code complexity)**	**Environment**
For any comments related to the task itself.	For any comments related to learning gains or language-specific comments.	For any comments that mention the environment itself (i.e. FTF or VW affordances.)
fun	authentic	fun
simple	simple	easier
difficult	difficult	WTC
	communication	technical
	pronunciation	affordances
	words (vocabulary items)	
	grammar	
	phrases	
	translation	
	listening	
	non-verbal	
	JP	
	ZPD	

For task-related comments, the main codes are “fun,” for when participants mentioned that they enjoyed the task and “simple” and “difficult” which were used when a participant specifically mentioned the difficulty of the task.

Language-related codes were used when a participant specifically mentioned learning or language-related aspects of the task. For example, the elements of the language they thought they had learned: words, phrases, grammar, pronunciation or communication in general. Other learning-related codes referred to how difficult or simple the language was, as well as whether they used their native language of Japanese (the “JP” code). Also, related to cooperation, some participants wrote that they were able to learn from their partner, which was coded as ZPD for the zone of proximal development.

Finally, comments that specifically mentioned modality (the VW or materials used during FTF communication) were labelled as “environment” as well as one of five subcategories. The code “easier” was only used when a participant mentioned that the current modality was easier than the alternative. WTC stood for willingness to communicate and was used for comments which mention that the particular mode promoted them to speak such as in the comment “*It was fun. I was able to use English effectively and didn't feel embarrassed. I learnt to express myself in English.*” Finally, the two codes “technical” and “affordances” can be considered opposites as “technical” was used with comments that mention any *technical difficulties* that participants encountered within that modality, and “affordances” refers to positive comments regarding the affordances of each mode. It was occasionally hard to distinguish whether a comment was referring to the environment’s affordances or the task’s simplicity or difficulty, in which case both codes were utilized. Examples can be seen in Table 25:

Table 25: Examples of environment-related comments

Modality	Environment-technical comment	Environment-affordances comment
FTF	The map could be made clearer. The stairs and bridges were difficult to see.	I think it is easier to explain something to my partner when I am face to face with them.
VW	I had difficulties speaking English and controlling the game.	It was hard to control but a lot of fun. Even if we make mistakes, we can remake the objects.

For open-ended question responses on the post-task comparison questionnaire, codes used were similar to those of the "environment" superordinate category above. However, as the post-task comparison questionnaire asked learners to compare the two tasks in a task pair, the superordinate categories used were "VW" or "FTF" (see Table 26). Codes "WTC," "affordances," and "technical" are as proposed above. "Cognitive-high" and "cognitive-low" were used when a participant specifically mentioned that a task was simple or difficult. Finally, "positive" was used when a comment mentioned that a task was fun or that they enjoyed the task. The subcategory "negative" was not added as a counterpart to positive because no comments specifically mention not enjoying a task.

Table 26: Codes used for open-ended question responses on the post-task comparison questionnaire

FTF	VW
WTC	WTC
affordances	affordances
technical	technical
cognitive-high (difficult)	cognitive-high (difficult)
cognitive-low (simple)	cognitive-low (simple)
positive	positive

3.5.2.1 Inter-coder reliability of post-questionnaire responses

Among the measures used in the post-task questionnaires, only the coding of open-ended responses required high inference. A native English speaker who is also a researcher in applied linguistics served as the second coder. The coder was briefed on how to code the

data in a 30-minute training session supervised by myself. I reviewed the coding guidelines verbally and provided some concrete coded examples to aid their understanding. During the subsequent coding session, the second coder coded all of the post-task questionnaire data. They were also encouraged to leave comments on responses that seemed difficult to code for a post-coding debriefing session, as well as provide any additional codes that seemed missing.

Inter-coder reliability was calculated to be 74.77%. This figure was calculated by dividing the total number of coded answers that were the same for both coders by the total number of coded answers:

$$\frac{Total\ number\ of\ coded\ answers\ that\ matched}{Total\ number\ of\ coded\ answers} = inter\text{-}coder\ reliability\ (\%)$$

Issues that came up in the coding system specifically related to the two codes "task-fun" and "task-difficulty" where the second rater would use "task-fun" in places where participants mentioned the difficulty of tasks. Such as in the response:

#2 難しかったから勉強になった。自分の英語能力をフルに活躍しないと相手に通じなかった。オンラインで学ぶには良い活動だと思う。I learnt a lot because it was a difficult activity. I had to use all of my English skills to explain myself to my partner. I think this activity is perfect for learning in virtual worlds. (Room VW individual post-task questionnaire)

The second coder coded this as "task-fun." Although the participant does not directly mention that the task was fun (in my interpretation), the second coder saw the fact that the participant "learnt a lot" because the activity was difficult had the underlying connotation that the task was fun. It was, therefore, necessary to decide what "task-fun" should refer to, and thus, update the coding schema as necessary to reflect this change.

Other problems arose where there were multiple interpretations of a particular response as in:

#3: 英語を話すのと、ゲームの操作で苦労した。 *I had difficulties with the English and game controls* (Room VW individual post-task questionnaire)

This response could be interpreted as the participant 1) having difficulties speaking

English and game controls as two separate issues, or 2) as having trouble speaking English whilst controlling the character, therefore a single issue brought about by the high cognitive demands of the task. In cases of such disputes, we explained our position and interpretation of the data, reaching an agreement about how codes should be applied, before moving onto the next issue in coding.

A final version of the coding produced by both coders was created based on the two interpretations, and the coding system was finalized. In light of the finalized version of the coding system, I went back over all of the data, reconsidering the initial codes with the new perspective gained from the second coder's opinions.

3.5.3 Statistical analyses

Before running statistical analyses on the data, normality checks were employed to check if the data were normally distributed, as one assumption to running an ANOVA test is that the data is reasonably normally distributed. Results of these checks showed that the data was normally distributed for the majority of measures, but not all. One common reason for data to appear non-normally distributed was identified.

Due to the participants differing English abilities, those students with the lowest and highest proficiencies were often marked as outliers based on their over- or underperformances, causing the data to appear skewed, or otherwise non-normally distributed. One higher-level participant, in particular, outperformed his peers in a number of measures as exemplified in the "different words" measure for the VW Room Decoration task (see Table 27, bolded text). These values are not outliers in the sense that they were incorrectly recorded, nor do they misrepresent those participants' performance during the tasks, and as such were kept in the data set. The decision to keep these outliers in the data was made upon the knowledge that the statistical test employed in this study, a two-way repeated measures ANOVA, is particularly robust to deviations from normality. However, in the case that outliers were discovered, the statistical analysis tests were run both with and without outliers included to ascertain if there was a statistically significant difference in results. If significant differences were found, the outliers were removed, and the outlier-free version of the test employed in further inspection instead. This method is considered one appropriate way of dealing with outliers in the statistics literature (Weisberg, 2014).

Table 27: Different words measure for the VW room decorating task.

Participant number	Different words used
1	52.00
2	157.00
3	46.00
4	39.00
5	86.00
6	45.00
7	142.00
8	254.00
9	108.50
10	98.00
11	9.00
12	84.50
13	57.50
14	132.00
15	100.00
16	120.00
17	123.00
18	129.00
19	97.50
20	116.00

3.5.3.1 Statistical tests employed in this study

Statistical tests used on the quantitative data in this study are: 1) descriptive statistics to generate means and standard deviations; 2) two-way repeated measures ANOVAs were used to explore the statistical significance of a two-way interaction between modality and task complexity on learner performance (analysis of spoken data) and perceptions of tasks (analysis of questionnaire responses); and finally 3) one-way repeated measures ANOVAs were used when exploring the simple main effect of task complexity on learners performance or perceptions. The alpha level for all statistical tests was set at $p < .05$. The software used

for all statistical analyses was IBM SPSS Statistics, version 24.

3.5.3.2 Two-way repeated measures ANOVA

The two-way repeated measures ANOVA was employed due to the design of the study. There are two within-subject factors explored in this study: modality and task complexity. There are two levels for modality and three levels for task complexity. Therefore this study can be considered a 2 x 3 factorial design. In order to run a repeated measures ANOVA, a number of assumptions must first be met (Lund & Lund, 2015):

1. **The dependent variable should be measured at the continuous level**.

"Continuous" means that there are no discreet groupings (ordinal or nominal variables). For the quantitative data generated from the CAF codes of the transcriptions, all CAF measures are recorded at the continuous level.

2. **The independent variable should consist of at least two related groups**.

There are two independent variables in this study: modality and task complexity and a single group of participants completed all factors, thus making this a repeated measures study. There are therefore three matching pairs or related groups of data.

3. **There should be no significant outliers**.

As mentioned above, statistical analyses are run with and without outliers. If any discrepancy in statistical significance is found between the two sets of data, this will be explicitly stated, and the data set used for further examination made clear.

4. **The distribution of the dependent variable should be approximately normally distributed.**

The data were initially examined for violations of the assumption of normality. This was done via skewness and kurtosis statistics, Shapiro-Wilk tests and by examining histograms, normal Q-Q plots and boxplots.

5. **The variances of the differences between all combinations of related groups must be equal (known as sphericity).**

The sphericity of data in this study is checked via Mauchly's test of sphericity, which

requires further explanation.

3.5.3.3 Mauchly's test of sphericity

Mauchly's test of sphericity was conducted to test the null hypothesis that the variance in differences between the levels of the within-subject factors are equal. In other words, in order to run a repeated measures ANOVA, there is an assumption that the variance in differences between all combinations of values for a specific group are equal. If the results of the test are statistically significant (i.e. if $p < .05$), sphericity is said to be *violated*, which means that the two-way repeated measures ANOVA is biased. The bias relates to how easily it returns a statistically significant result. There is a way to correct for this bias, which involves adjusting the degrees of freedom used in calculating significance (or the "p" value). The correction is known as epsilon (**ε**), and the Greenhouse-Geisser method is used to generate it.

If sphericity is assumed, however, this indicates that the two-way repeated measures ANOVA is not biased, and no adjustments need to be carried out on the degrees of freedom. For instance, consider the following fictitious data in Table 28. The variance appears unequal (13.9 vs. 17.4 vs. 3.1) but running Mauchly's test of sphericity reveals that the data does not violate the assumption of sphericity (p =.19).

Table 28: An example of sphericity.

Participant	Task 1 utterances	Task 2 utterances	Task 3 utterances	Task 1 – Task 2	Task 1 – Task 3	Task 2 – Task 3
1	45	50	55	-5	-10	-5
2	42	42	45	0	-3	-3
3	36	41	43	-5	-7	-2
4	39	35	40	4	-1	-5
5	51	55	59	-4	-8	-4
6	44	49	56	-5	-12	-7
			Variance:	13.9	17.4	3.1

In summary, based on the results of the sphericity test, the significance value of any results (p) is either taken as is (sphericity assumed), or adjusted via the Greenhouse-Geisser

method (sphericity violated).

3.5.3.4 Identifying interaction effects

This section introduces the steps followed in interpreting the results of the ANOVA tests. Initially, the results are analysed to determine whether there is a statistically significant two-way interaction between the two within-subject factors (i.e. interaction between *modality* and *task complexity*). If a statistically significant two-way interaction is found, the results are analysed further to determine if there are any simple main effects. However, if a two-way interaction is not found, the results are analysed for statistically significant main effects. Figure 13 shows a summary of how the data was interpreted.

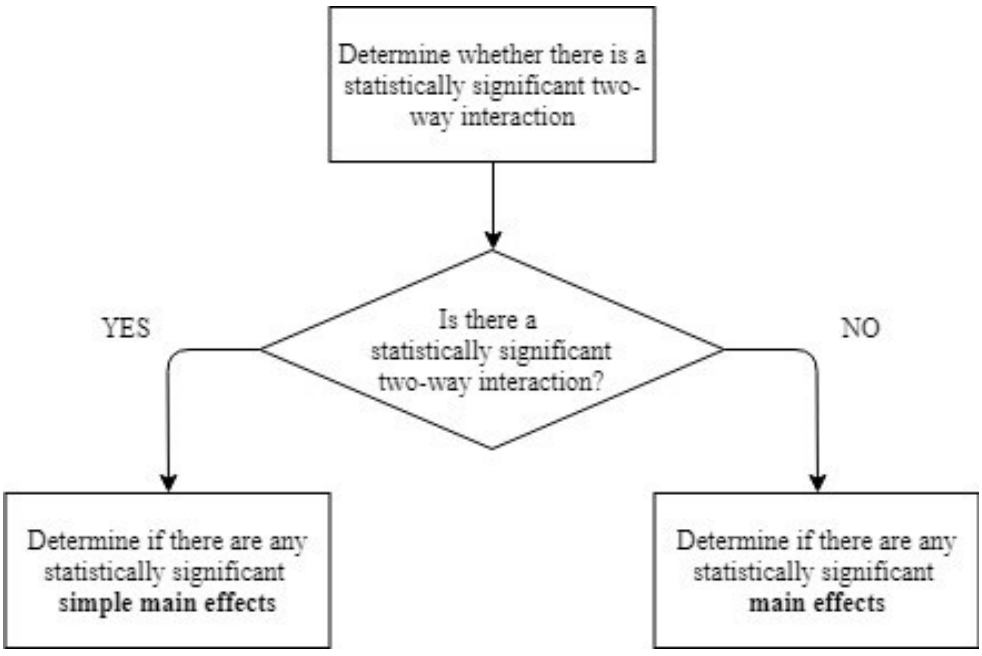


Figure 13: Flowchart for interpreting results of two-way repeated measures ANOVA tests.

A two-way interaction was determined in the following manner: First, profile plots produced by the test were visually inspected to verify any trends. One holistic approach to identifying whether an interaction exists is by considering how parallel the two lines are. Subsequently, Mauchly's sphericity test was employed to establish if the assumption of sphericity had been violated. Finally, if sphericity was assumed, tests of within-subjects effects were referenced. Specifically, the "mode * task complexity" box. Table 29 shows the results of a two-way repeated measures test on the variable "syllables per utterance. The results of this test do not reveal a statistically significant two-way interaction, $F(2, 38) =$

4.00, $p = .17$.

Table 29: Sample data which does not show a statistically significant two-way interaction.

Source		SS	df	Mean Square	F	Sig.	Partial Eta Squared
Mode * Task Complexity	Sphericity Assumed	2.7	2	1.33	4.00	.03	.17
	Greenhouse-Geisser	2.65	1.75	1.51	4.00	.03	.17
Error (Mode* Task Complexity)	Sphericity Assumed	12.62	38	.33			
	Greenhouse-Geisser	12.62	33.31	.38			

3.5.3.4.1 Interpreting simple main effects

Upon finding a statistically significant two-way interaction, the data was further investigated for **simple main effects**. The reason for exploring simple main effects is to understand at what level the independent variables had a significant effect after an interaction has been found. Simple main effects are the differences between mean cell scores and can be initiated at any level of the two factors employed in the tests (in the case of this study: giving preference to either modality or task complexity). In this way, the results of the two-way repeated measures ANOVA can be used to answer Research Question 1 with a focus on both modality and task complexity or the two factors separately depending on how simple main effects are investigated. As a concrete example, simple main effects could be determined by looking for significant differences in the mean scores of the number of syllables per utterance measures when participants completed the FTF or VW tasks, thus providing an answer to Research Question 1 in terms of learners' output complexity between modes of communication (see Figure 14 for an example). SPSS does not have the functionality to run simple main effects, but it is possible by using the *GLM: Repeated Measures* procedure on subsets of the data. In other words, simple main effects can be investigated by running separate one-way repeated measures ANOVAs on the data. As there three levels of the within-subjects factor for modality in this example, three separate tests are required (red boxes).

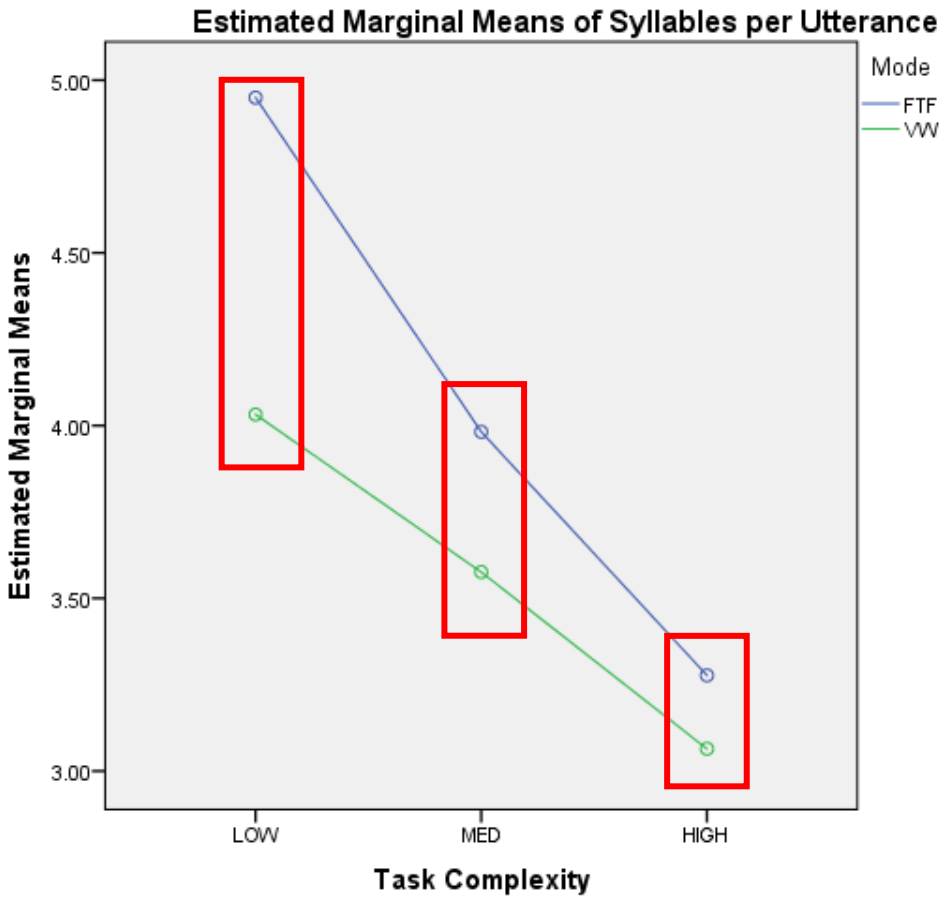


Figure 14: Example of testing for a simple main effect between modes of communication.

In determining the simple main effects for **modality,** only two levels are compared, and so there is no need to check from the assumption of sphericity. Referencing the "test of within-subjects effects" and "pairwise comparisons" tables reveals any simple main effects. However the same is not true of **task complexity**. With a comparison of three levels (low, mid and high task complexity), the assumption of sphericity must be considered, and there will be multiple additional pairwise comparisons. Figure 15 outlines this graphically.

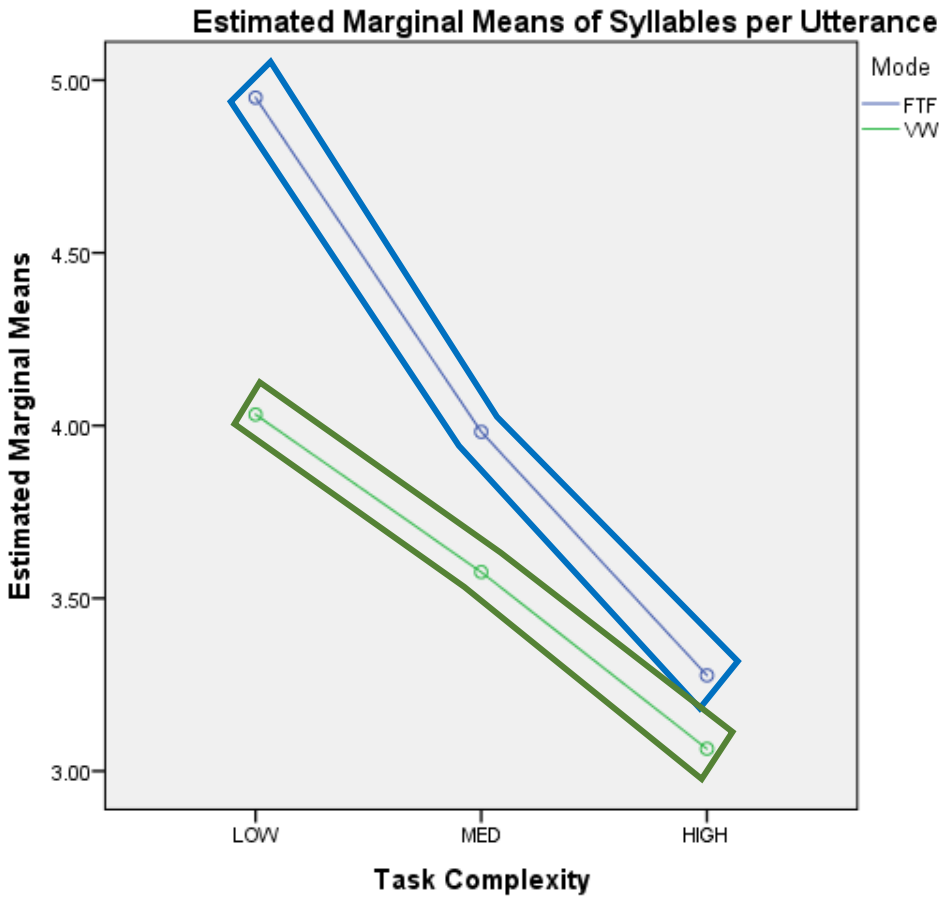


Figure 15: Example of testing for a simple main effect for task complexity.

3.5.3.4.2 Interpreting main effects (procedure for no significant two-way interaction)

If there is no statistically significant two-way interaction found between modality and task complexity, the **main effects** for the two within-subject factors are interpreted. The main effect for task complexity is calculated by comparing the mean scores of LOW, MID and HIGH complexity tasks. The main effect of modality is calculated by comparing the mean scores of FTF and VW tasks.

Similar to the steps followed above for investigating the simple main effects, the main effect for modality has only two levels, and so the assumption of sphericity is not required. The result of the main effect for modality can be found in the **tests of within-subjects effect** table of the two-way repeated measures ANOVA test in SPSS, which is produced when investigating whether there is a two-way interaction. Again, similar to the simple main effects investigation, for task complexity, the main effects for task complexity are investigated by

referring to the same tests of within-subjects effects table.

3.6 RELIABILITY AND VALIDITY CONSIDERATIONS

3.6.1 Reliability

The reliability of a study relates to how replicable the results are. This is revealed in two ways: *internal* and *external* reliability. A study is said to have internal reliability if an independent researcher would come to the same conclusion upon reanalysis of the data. A study would have external reliability if an independent researcher reached the same conclusion by replicating the study.

The current study is conducted from a generally positivist research paradigm, that is, the analysis of quantitative data, and as such an independent researcher should reach the same conclusions if using the same data set and data analysis techniques. The reliability of the CAF measurements has been established based on their adoption in previous studies (for example see Ellis & Yuan, 2003; Foster, Tonkyn and Wigglesworth, 2000; Foster & Skehan, 2013). However, the transcription CAF coding techniques employed as the primary source of data collection for this study do prompt a cause for concern in terms of internal reliability. That is, unless the coding system is outlined in detail, it would be impossible for the study to be replicated. Therefore a detailed description of this system was provided above (section 3.5.1). Additionally, spelling mistakes and other input errors made during the transcription process present another cause for concern. The spoken discourse of the students was transcribed by the author and then initially put through computer spellcheck to find any errors. After this stage, the audio was listened to again in parallel to reading the transcribed data to check for any input errors, thus increasing the internal reliability of the study.

The second issue is the coding the transcription data in terms of the CAF measures (errors, number of syllables, clauses, etc.). A drift in coding definitions can occur if the coding schema is not rigorously defined at the outset. In the case of the current study, initially, a sample section of data was selected, coded, and put aside for two weeks. After this period, the same data was recoded. Upon completion, the two pieces were cross-referenced for any inconsistencies. This method helped to reveal whether the coding system was replicable. Inconsistencies found were due to careless errors rather than coding definition discrepancies. Thus, in order to reduce the number of errors in coding, all data were rechecked between two

weeks and one month after the initial coding session. If any discrepancies were found, the data would be adjusted accordingly. Additionally, the coding schema was written down as a simple bullet point list to allow for quick reference during the coding sessions.

Finally, as outlined in detail in Section 3.5.2.1, a coding schema for the post-task questionnaires was created, and then subjected to a second coding session by another coder where inter-coder reliability was found to be 74.77%. The action of getting a second coder's opinion on the data helped discover weaknesses in the original coding schema which was subsequently updated. The data was then recoded with reference to the finalized coding schema, helping to increase the reliability of the study.

3.6.2 Validity

In terms of validity, Cohen et al. (2000) state, 'internal validity seeks to demonstrate that the explanation of a particular event, issues or set of data which a piece of research provides can actually be sustained by the data' (p. 107). This study is concerned with whether modality and task complexity has an effect on learners' oral task performance. In this way, a number of task pairs were created in order to increase the possibility that measures of task performance are accurately interpreted. In other words, instead of utilizing a single task pair to explore the effect of modality on task performance, three different task pairs allow for the collection of a larger data set, and therefore more internal validity checks. For instance, if all three VW tasks result in the accuracy of learner performance to be lower than their FTF counterparts, by utilizing pairwise comparisons of statistical analyses, it is possible to make the claim that modality negatively affected accuracy. Having more than one set of tasks has benefits in testing the claims of the Cognition Hypothesis, and, more specifically, the effect of task complexity on learner task performance also. Tasks were also trialled in a pilot study to ensure the appropriateness of task design (for example the discussion on open and closed-goal orientation), as well as pre- and post-task materials.

Task complexity variables were operationalized to keep them as similar as possible between tasks in a task pair (such as the number of characters to find in the spot-the-different tasks). The same is true of *task conditions* and *task difficulty* variables also. For example, care was taken to make sure that tasks were sequenced in a counterbalanced design within each instance of the class, and between instances of the class (i.e. tasks were sequenced such

that the order of task completion was different between task pairs and between classes).

Dyad creation was also carefully considered, students that knew each other outside the class were separated from one another, and was paired with an unknown partner to keep participant-related task conditions as equal as possible. In terms of the participants themselves, a background check on their experiences with games and technology, as well as their English proficiency were all recorded at the start of the course, and the data used as covariates in statistical analyses to ascertain whether such proficiencies had an influence on their task performance. The results of these tests did not reveal any statistically significant differences in mean scores however and are thus not reported in this study.

One positive element of this study in terms of its internal validity is that it adopts a repeated measures design, meaning that all subjects start at the same point (having not done any of the tasks) and their performances are measured over the course of the experiment. Individual differences are therefore significantly reduced in such designs, and changes in performance are therefore more easily attributed to the effect of independent variables. The disadvantage to such studies, however, is the order effect, in other words, the order of tasks could have a more significant effect on performance than any one individual task, which is related to the Resultative Hypothesis of task repetition (Skehan, 1991). However, as described above, this was considered, and the study adopted a counterbalanced designed.

External validity may be considered the weakest part of the study in that it refers to how generalizable any findings are to the broader population. For instance, convenience sampling techniques were employed in this study, which means that the sample population was comprised of a homogenous group of low-level English learners with similar educational backgrounds. Additionally, the majority of participants were male, and of a similar age range. Another concern is the number of participants in the sample of this study. As a rule of thumb, the minimum sample size that is required is twenty subjects per group (Marion, 2004), and this study only just meets that requirement. These limitations thus acting as prompts for further studies within this field of inquiry.

3.7 ETHICS AND INFORMED CONSENT

As part of any study, it is important to consider both appropriate research methods and ethical considerations. This is especially true of social sciences where the subjects of

research are often human beings. The current study was informed by the six fundamental principles of the Economic and Social Research Council's (ESRC) research ethics framework (2015). These are:

- research should aim to maximise the benefit for individuals and society and minimise risk and harm
- the rights and dignity of participants should be respected
- participation should be voluntary and involved informed consent
- research should be conducted with integrity and transparency
- responsibility and accountability should be clearly defined
- independence of research should be clear, and conflicts of interest should be made explicit.

Of course, the assumptions and implications of these six principles have been criticised in a number of papers (see Hammersley, 2010), however for the purposes of the current study, due to their level of simplicity in implementation and relative lack of invasiveness, they were deemed appropriate. These points informed the code of conduct for the current study with the aim of upholding the strong research practices of the University of Leicester and maintain that any results of the study were not invalidated by a lack of consideration of critical ethical issues.

In order to conduct this research, I had to seek approval from the University of Leicester based on an internal review procedure. Approval was granted based on the inclusion of a consent form which was distributed and completed by the participants at the start of the course (Appendix 1). This form included information on the study, the right of the participant to withdraw at any time and the confidentiality and anonymity of the participant in the written report. As it is essential to match questionnaire data with the spoken data, participants needed to provide me with their student number as part of the data collection process. However, this identification number does not appear in the final research report, thus protecting the anonymity of participants (Bryman, 2015)

Audio data was collected from the participants' PC terminals via a USB memory stick and then stored on my personal computer in my office. This data was then transferred to an online storage site, which again was only accessible to me via a secure password. All paper-

based data were stored in a locked cabinet in my office.

4 RESULTS

This chapter reports on descriptive and statistical analyses of the data collected in this study. The data was collected as part of two separate interventions. Due to the elective nature of the class, the two instances used in this study were comprised of a different number of students. The first instance consisted of eight students, one female and seven males. The second class was comprised of 20 male students, although not all of them participated in the study. A total of 18 students (9 dyads) agreed to participate, however, absences and technical issues with the VW recordings for a number of the dyads meant that data was only collected from a total of 12 students (6 dyads). See Table 30 for an overview.

Table 30: Breakdown of participant numbers.

Instance		**Number of students**	**Dyads**
First			
	Total	8	4
	Initial participants	8	4
	Unusable data	N/A	
	Final participants	8	4
Second			
	Total	20	10
	Initial participants	18	9
	Unusable data	5	4
	Final participants	12	6
Total		**20**	**10**

Regarding technical issues, the project met with issues throughout its undertaking. One major issue was with the hardware in the classroom. Multiple times during the second intervention, a switcher ceased to supply an internet connection to four PCs, which meant that those participants had to be reseated. Fortunately, this issue was discovered upon initial entering of the VW, and so no loss of recorded data occurred.

Recording students as they undertook the VW tasks also caused several problems. As mentioned, *TeamSpeak* was the software used for mediating oral communication between

participants. It was also used to record their voices during the VW tasks. This software utilizes a push-to-talk mechanism much like a two-way radio transceiver: interlocutors can only hear you speak when you are holding down a specific key (the control key by default). The recording is only activated when this key is held down also. The issue was that participants would forget to press the key down. This prompted their interlocutor to either rephrase their statement or ask a question again, as they believed that they had not been heard correctly. This occurrence was infrequent in the recordings, but still a cause for concern. It may have had a negative effect on participants' output complexity and fluency, as there were either extended stretches of silence or repetitious use language of language, reducing MTLD scores.

Regarding participant-related issues, absences were the most significant cause for concern and were the cause of losing a substantial amount of data. Another major issue related to the layout of the room and again, the elective nature of the class. Because the room was so large and participants were so widely dispersed, it was difficult for me to listen in on participants as they completed tasks. Transcribing the recordings of one particular group, I found that they had completed the first set of tasks using English, and so the transcription was created. However, opening their file for the second task pair, I found that they completed the task predominantly using the L1 (Japanese). This data was therefore considered unusable, and data for the pair's other tasks (as well as questionnaire data) were removed from the study. Reasons for why they chose to do the task in Japanese was not explored further, however, assumptions are that either the task was too cognitively difficult to complete in English (their proficiency level was too low), or, due to having low motivation towards studying English, they were not invested in completing the class activities in English.

4.1 VALIDATING TASK COMPLEXITY PREDICTIONS

Three levels of task complexity were created (LOW, MID and HIGH complexity), based on the operationalization of task conditions (+/-) number of elements, (+/-) spatial-reasoning, and (+/-) negotiation needed. In terms of modality, it was hypothesised that the VW tasks would be more complex than their FTF equivalents due to the additional cognitive demands of technology mediating communication, increased spatial reasoning requirements, and technical skills required to perform the tasks in the VW.

The post-task questionnaires are first analysed in order to determine if task complexity was perceived by participants as predicted, a crucial step in validating task complexity manipulations (Révész, Michel, & Gilabert, 2016; Sasayama, 2015). That is, in determining whether learners' perceptions of task complexity match predictions, it then is possible to investigate how task complexity affected oral task performance.

4.1.1 Post-task questionnaire quantitative data

Mean scores and standard deviations were calculated for each measure on the post-task questionnaire and can be seen in Table 31 below. In this section, answers to each question are explored in detail. Two-way repeated measures ANOVAs were performed on the data with a focus on uncovering the existence of a two-way interaction between modality and task complexity on participant perceptions of task complexity for each measure.

The first four questions relate specifically to how modality and task complexity influenced participants' task difficulty perceptions, thus helping to validate task complexity predictions and answer RQ1 (i.e. the effect of modality and task complexity on learner output). The fifth question relates to the effect of modality and task complexity on participants enjoyment of the tasks. Findings gained from this measure are used to answer RQ2 (i.e. the affective affordances of the VW).

Table 31: Means (and Standard Deviations) for each measure on the post-task questionnaires.

Task	Mental Effort	Task Difficulty	Difficulty of vocabulary	Attentional focus*	Enjoyment
LOW FTF	3.35 (0.93)	2.65 (0.81)	2.40 (0.82)	2.60 (1.19)	4.05 (0.60)
LOW VW	3.50 (1.05)	2.90 (0.91)	2.20 (0.77)	2.70 (1.42)	4.55 (0.60)
MID FTF	4.00 (0.79)	3.75 (0.85)	3.00 (1.03)	2.00 (0.73)	4.10 (0.64)
MID VW	4.70 (0.73)	4.85 (0.37)	4.10 (0.72	1.75 (1.25)	4.20 (0.77)
HIGH FTF	3.50 (1.15)	2.80 (1.06)	2.25 (0.85)	1.75 (0.64)	4.15 (0.59)
HIGH VW	3.95 (0.89)	3.75 (0.79)	2.45 (0.89)	2.35 (1.09)	4.40 (0.50)

Note. *Attentional focus is ranked 1 (motor skills) – 5 (speaking English)

The questionnaire data were examined for outliers first. Three were found which had studentized residual values of -3.41 (VW MID Mental Effort), -3.78 (VW HIGH, Mental

Effort) and 3.32 (FTF MID Vocabulary Difficulty). Subsequently, the data were further examined for violations of the assumption of normality. The data was not normally distributed for any of the measures as assessed by Shapiro-Wilk's test of normality on the studentized residuals. However, an inspection of Q-Q plots revealed that apart from the three outliers found above, the data were reasonably normally distributed. Thus, the data were not transformed. Mauchly's test of sphericity indicated that the assumption of sphericity was met for a two-way interaction between modality and task complexity for all measures. However, only two of the measures exhibited a statistically significant interaction between modality and task complexity: *Vocabulary Difficulty* ($F(2, 38) = 6.45$, $p = .004$) and *Task Difficulty* ($F(2, 38) = 3.93$, $p = .03$). There was no statistically significant two-way interaction between modality and task complexity for the measures *Enjoyment* ($F(2, 38) = 2.65$, $p = .08$), *Focus* ($F(2, 38) = 1.76$, $p = .19$), or *Mental Effort* ($F(2, 38) = 1.62$, $p = .21$).

The following sections, therefore, introduce the results of the statistical analyses conducted on the quantitative data in two separate sections: Simple main effects are investigated for the *Vocabulary Difficulty* and *Task difficulty* measures first, and, following that, main effects are investigated for the remaining three measures.

4.1.1.1 Simple main effects of modality and task complexity on perceptions of vocabulary difficulty

This section introduces the simple main effects of modality and task complexity on participants' perceptions of the difficulty of encountered vocabulary during all six tasks. A visual plot of mean scores is available in Figure 16.

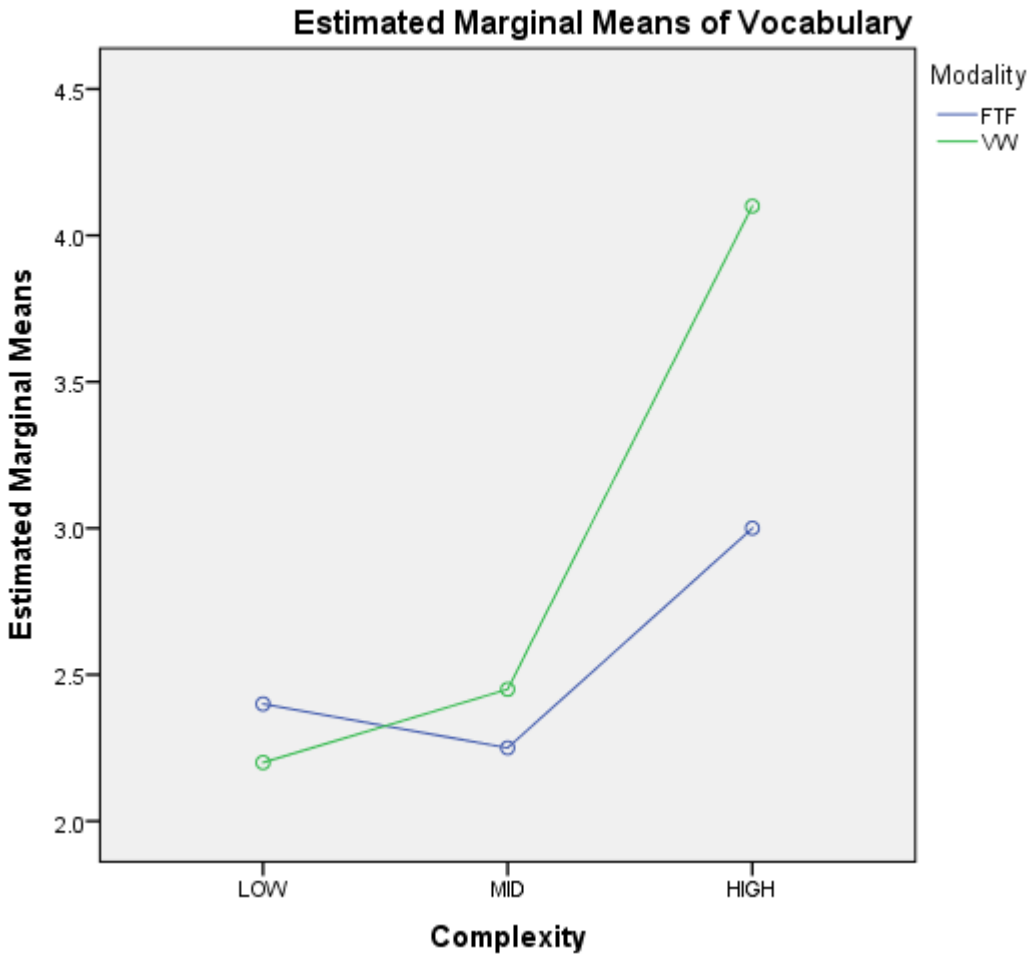


Figure 16: Mean vocabulary difficulty scores across task complexity

In terms of modality, there was only a statistically significant difference in mean scores at the HIGH complexity task level, $F(1, 19) = 14.46$, $p = .001$, where mean scores were 1.1 points higher for the VW modality. There was no statistically significant difference in mean scores at the MID and LOW levels. Subsequently, there was no statistically significant effect of task complexity on participants' perception of vocabulary difficulty in the FTF mode, $F(2, 38) = 3.35$, $p = .08$, however, there was for the VW mode, $F(2, 38) = 37.91$, $p < .001$. Examining pairwise comparisons for the VW tasks further revealed a statistically significant difference in mean scores at the LOW-HIGH and MID-HIGH levels, where the HIGH complexity mean score was 1.9 points higher than the LOW complexity task, $p < .001$, and 1.65 points higher than the MID complexity task, $p < .001$.

In summary, the vocabulary required for successful completion of the HIGH

complexity tasks was considered more difficult than the LOW and MID complexity tasks, and in addition, the VW HIGH task was considered to have the most difficult vocabulary requirements of all six tasks, a finding somewhat consistent with task complexity predictions.

4.1.1.2 Simple main effects of modality and task complexity on task difficulty perceptions

A visual plot of mean scores for the task difficulty measure is provided in Figure 17. Modality had a statistically significant effect on participants perception of task difficulty at the MID ($F(1, 19) = 18.10$, $p < .001$), and HIGH ($F(1, 19) = 29.10$, $p < .001$) complexity levels. The mean score for the VW MID task was 0.95 points higher than the FTF equivalent, and the VW HIGH task was 1.1 points higher than the FTF equivalent. In general, then, the VW tasks were considered more cognitively demanding than the FTF tasks, but only statistically significantly so at the MID and HIGH levels.

For simple main effects of task complexity in the FTF mode, there was a statistically significant difference found in mean scores at the LOW-HIGH and MID-HIGH levels, where the mean score for the HIGH complexity task was 0.95 points higher than the MID, $p = .01$, and 1.1 points higher than the LOW complexity tasks, $p < .001$. Additionally, for the VW mode, task difficulty perceptions were statistically significantly different at all three levels. Findings echo those of the previous measure in that whilst at LOW task complexity levels there is little effect of modality or task complexity on participants' perception of task difficulty, as task complexity increases, there is an interaction between modality and task complexity, which results in the VW HIGH complexity task being perceived as particularly cognitively demanding.

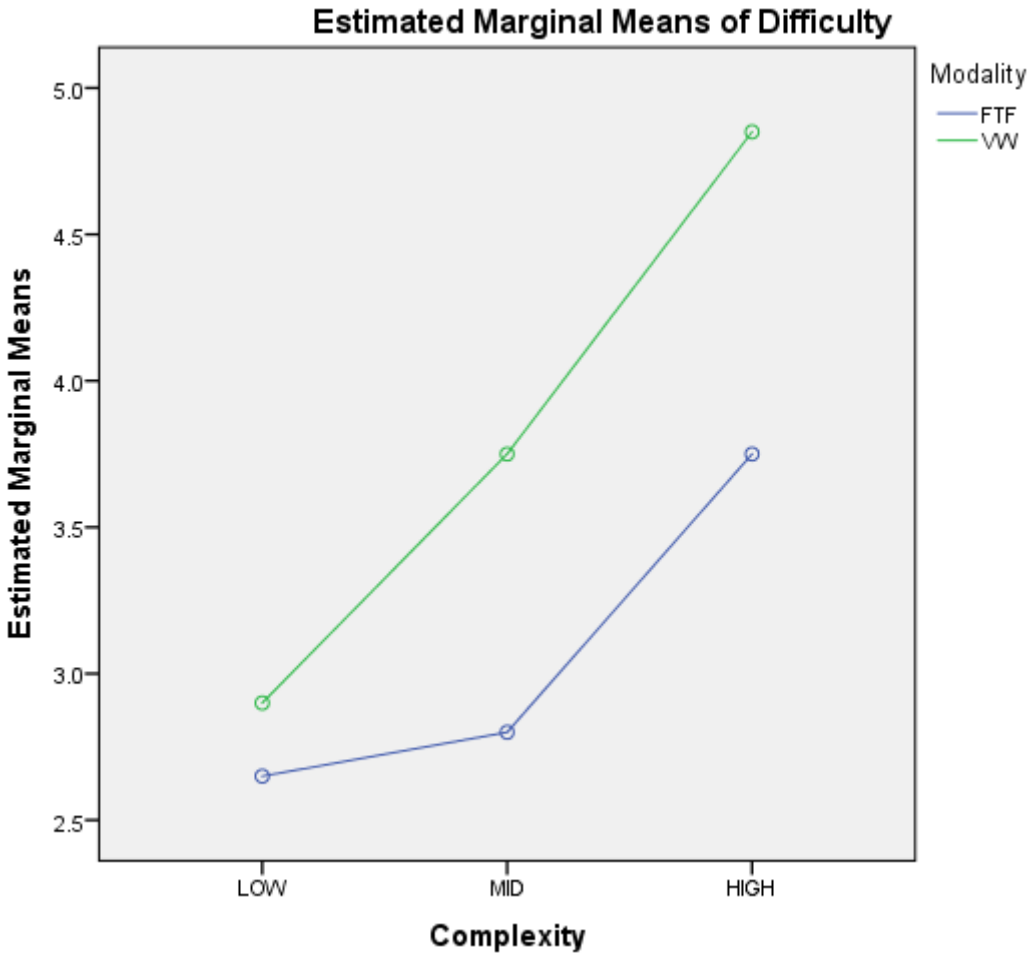


Figure 17: Mean task difficulty scores across task complexity

4.1.1.3 Main effects of modality and task complexity on task enjoyment, attentional focus and mental effort.

There was no statistically significant interaction effect between modality and task complexity for three measures on the post-task questionnaire. Therefore the main effects of modality and task complexity were investigated instead of simple main effects.

To recall, a high score for the *focus* measure indicates that participants focused more attention on language production, a low score indicates that they focused more attention on motor skills. A low score, therefore, suggests that the cognitive demands of a task were too high for them to focus on language production. A visual plot for this measure is available in Figure 18. There was no statistically significant main effect of modality on attentional focus, $F(1, 19) = 0.62$, $p = .44$. However, there was task complexity, $F(2, 38) = 5.92$, $p = .006$. Post hoc analysis with a Bonferroni adjustment revealed that there was a statistically significant increase in a focus on language production when participants completed the LOW

complexity tasks compared to the HIGH complexity tasks (mean difference = 0.76, $p = .01$). Statistically significant differences were not found at the LOW-MID and MID-HIGH levels. Thus, regarding participants focus of attention, modality had no main effect. However, there was a slight tendency for the lower complexity tasks to free up attentional resources for language production, which suggests that task complexity manipulations were as predicted – increasing task complexity hindered participants' ability to focus attention on language production.

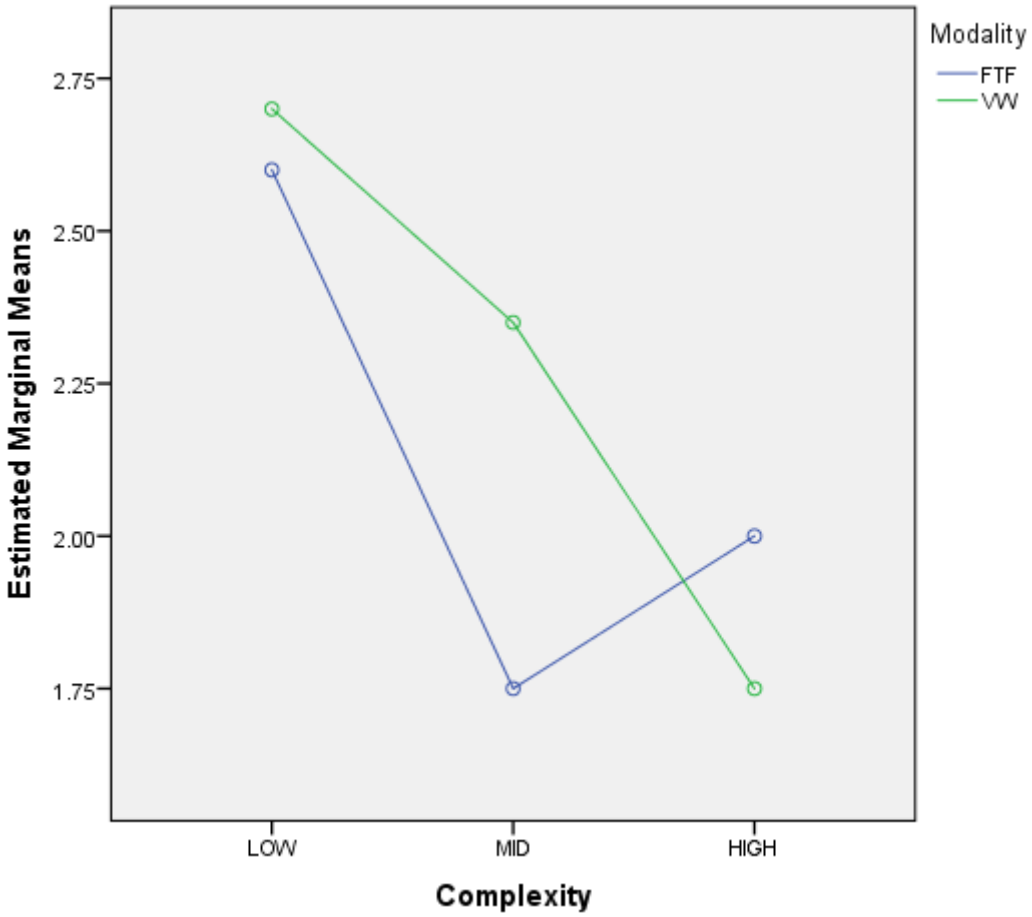


Figure 18: Mean scores for focus across task complexity

A visual plot of mean scores for the *mental effort* measure is available in Figure 19. Statistical analysis revealed that there was a main effect of modality on mental effort where the mean score for the VW tasks was .43 points higher than the FTF equivalents, $F(1, 19) = 15.22$, $p = .001$. There was a main effect of task complexity on participants' perceived level of mental effort also, $F(2, 38) = 10.97$, $p < .001$. Post hoc analysis with a Bonferroni

adjustment revealed that there was a statistically significant increase of 0.63 between mean scores of the MID and HIGH complexity tasks, $p = .008$, and an increase of 0.93 between the LOW and HIGH complexity tasks, $p = .001$.

Results for the mental effort measure thus suggest that the VW posed a higher level of cognitive demand than the FTF tasks, and there was a positive correlation between increasing mental effort with task complexity from LOW to HIGH as expected.

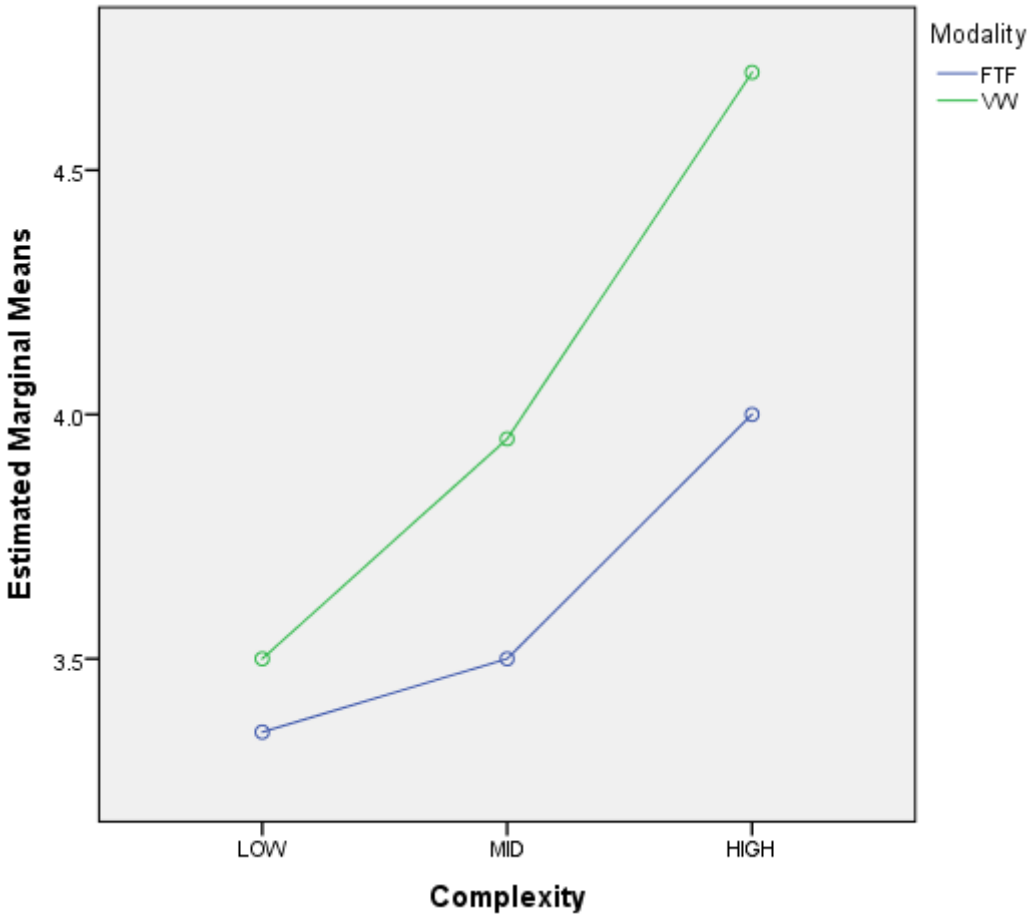


Figure 19: Mean mental effort scores across task complexity

Finally is the *enjoyment* measure which asked participants to rate their enjoyment of the six tasks (Figure 20). A higher score indicates greater enjoyment. There was a statistically significant main effect for modality, where mean scores for the VW tasks were 0.28 points higher than the FTF tasks, $F(1, 19) = 7.12$, $p = 0.15$. The main effect of task complexity showed no statistically significant difference in mean scores for participants' task enjoyment, $F(2, 38) = 1.54$, $p = .23$. Whilst this measure is not directly related to learners' perception of

task difficulty, the strong correlation between modality and task enjoyment is important in answering RQ2. Additionally, of particular note is that although the VW tasks were perceived to be more cognitively demanding than the FTF tasks, they were also considered to be more enjoyable. The positive relationship between cognitive demand, modality and enjoyment implies that either the VW tasks were intrinsically more interesting and thus enjoyable, or that the more complex task in a task pair was enjoyable *because* of the inherent challenge it posed to participants.

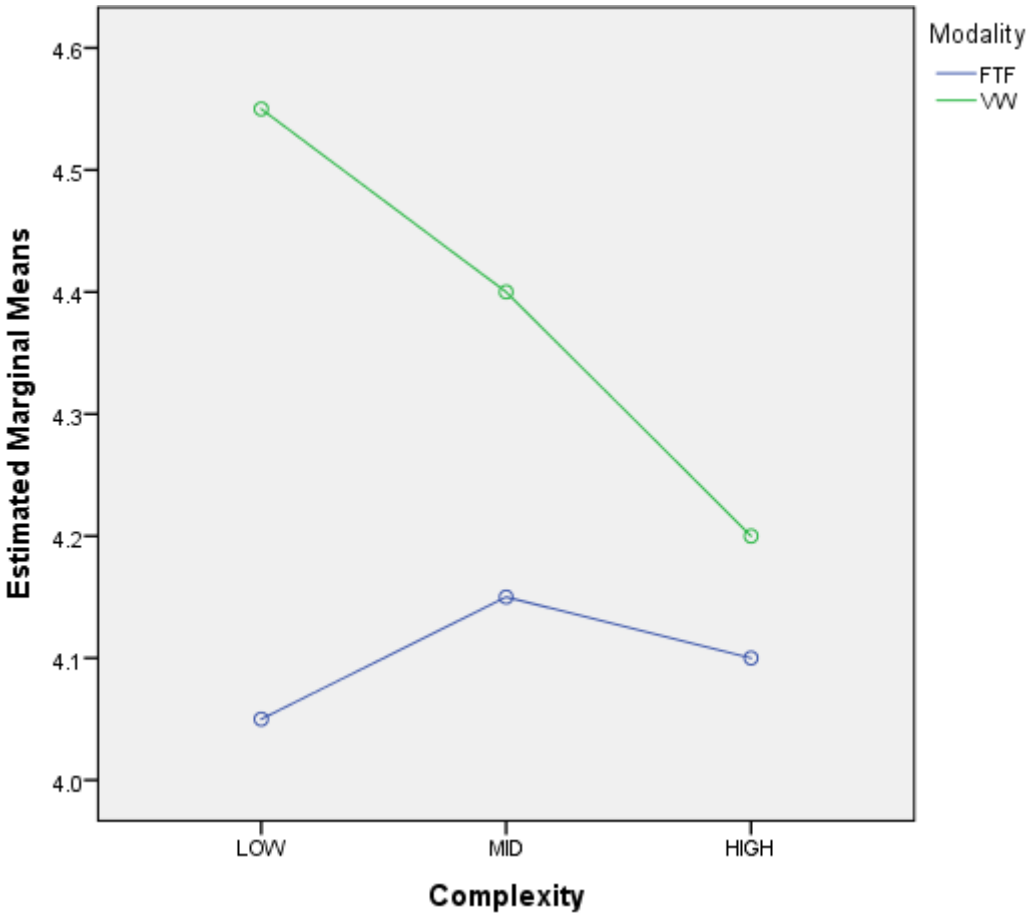


Figure 20: Mean enjoyment scores across task complexity

4.1.1.4 Summary

Regarding the post-task questionnaire, results suggest that, as expected, the VW tasks were considered more cognitively demanding than their FTF equivalents, requiring learners to exert more mental effort. Task complexity also had a significant effect on perceived task difficulty, where the HIGH complexity tasks were considered to be of higher difficulty than

the LOW complexity tasks. This result is also consistent with task complexity predictions.

In terms of their attentional focus, modality did not appear to affect performance, whereas task complexity did. This is a slightly surprising result, as it was predicted that the unfamiliarity of VW-based communication would demand more of participants' attentional resources thus be perceived as requiring more focus than the FTF equivalents. However, the independent variable task complexity did have an effect on attentional focus, where participants were able to focus more of their attentional resources on language production when completing the LOW complexity tasks compared to the HIGH complexity tasks.

For the measure regarding the difficulty of vocabulary encountered in each task, modality only had a significant effect on participants' perceptions at the HIGH complexity level. This may suggest that there was a closer balance in terms of vocabulary required to complete the LOW and MID tasks across modality, whereas, at the HIGH complexity level, the vocabulary requirements of the VW task were perceived as much more difficult and therefore suggesting a possible imbalance in task design. Alternatively, it could be that the VW HIGH task pushed learners to use more vocabulary due to the affordances of the mode. The highest mean score for the *vocabulary difficulty* measure was recorded for the Room VW task, which suggests that this task, in particular, may have required learners to use the largest volume of different words. The output complexity measure the *number of different words* may confirm participants' perception of this task pushing them to use the highest number of different words.

Finally, regarding the *enjoyment* of the tasks, all recorded mean scores were above the median value (= 3) suggesting that the tasks were generally considered to be enjoyable. VW tasks were considered more enjoyable than the FTF equivalents suggesting that although the VW tasks were considered both more mentally demanding and difficult than the FTF tasks, it was these tasks that were more enjoyable. Reasons for this are explored in more detail below, with reference to qualitative data recorded in the open-ended questions section of the questionnaires.

4.1.2 Post-task comparison questionnaire Likert-like question data

As well as completing the individual post-task questionnaires upon completing each task, participants answered an additional questionnaire after completing both tasks in a task

pair. The post-task comparison questionnaire was designed explicitly for participants to compare their performances and perceptions of the two tasks in a task pair (thus comparing the effect of modality on the four measures). Four Likert-like items appeared on the questionnaire asking them to compare spoken output, difficulty, enjoyment, and potential learning gains for the two tasks. The statements were weighted from 1 – 5. An answer of 1 showed a preference for the VW tasks and an answer of 5 for the FTF tasks. Following, descriptive and statistical analyses of the data are presented (Table 32).

Table 32: Means (and Standard Deviations) for each measure on the post-task questionnaires.

Task complexity	Pushed output	Task Difficulty	Language learning potential	Enjoyment
LOW	3.00 (0.97)	2.40 (1.05)	2.70 (0.98)	1.90 (0.79)
MID	3.25 (1.09)	3.05 (1.16)	3.00 (1.10)	2.00 (1.05)
HIGH	3.30 (1.45)	1.15 (0.49)	1.75 (0.91)	2.45 (1.23)

Note. 1 = virtual world preference, 5 = face-to-face preference

Whilst the majority of scores appear close to the median value of 3, there is a trend for the data relating to difficulty and enjoyment to be lower than this value. This indicates 1) that the VW tasks were considered more difficult than the FTF equivalents (a finding that appears above with the post-task questionnaires), 2) that the VW tasks were more enjoyable than the FTF equivalents (again, found above), and 3) that there is a correlation between task difficulty and enjoyment. One-sample t-tests were run on the data to test whether the mean score for each measure was statistically different from the median value. Results are provided in Table 33 with statistically significant results bolded.

Table 33: One sample t-tests on all questionnaire data

	t	Sig. (2-tailed)	Mean Difference	95% CI	
				Lower	Upper
Pushed Output LOW	.00	1.00	.00	-.46	.46
Pushed Output MID	1.00	.33	.25	-.27	.77
Pushed Output HIGH	.92	.37	.30	-.38	.98
Difficult LOW	-2.57	.02*	-.60	-1.09	-.11
Difficult MID	.19	.85	.05	-.51	.61
Difficult HIGH	-16.90	.000*	-1.85	-2.08	-1.62
Enjoyable LOW	-6.24	.000*	-1.10	-1.47	-.73
Enjoyable MID	-4.16	.001*	-1.00	-1.50	-.50
Enjoyable HIGH	-1.99	.06	-.55	-1.13	.03
Learning Potential LOW	-1.37	.19	-.30	-.76	.16
Learning Potential MID	.00	1.00	.00	-.53	.53
Learning Potential HIGH	-6.14	.000*	-1.25	-1.68	-.82

Note: Test value was set at 3.

*The mean difference is significant at the .05 level.

Mean scores for the first measure regarding which modality promoted learners to produce more output were not weighted towards either the FTF or VW mode, and results of the t-test revealed no statistically significant difference from the median score of 3, indicating that participants did not perceive modality as influential in pushing output.

For task difficulty, the LOW and HIGH task mean scores are statistically significantly lower than the average of 3. Modality at these levels, therefore, influenced perceptions of task difficulty, where the VW tasks were considered more difficult. The same is not true at the MID complexity level, where the mean score is not statistically significantly lower than the median score of 3.

Mean scores for the learning potential of tasks are similar to those of the task

difficulty measure, where both LOW and HIGH levels are weighted towards the VW tasks, however only the HIGH-level tasks were statistically significantly lower than the median score of 3. The difficulty of the HIGH-level VW task, therefore, may have influenced learners perception of its potential for learning. For the MID level tasks, the mean is not statistically different from the neutral score, indicating that neither the FTF or VW task was perceived as having more potential for learning at this level.

Finally, mean scores for enjoyment indicate that the VW tasks were more enjoyable at all complexity levels. However, only two of the mean scores were statistically lower than the median score (LOW and MID level tasks). This finding corroborates somewhat with the post-task questionnaire measure for enjoyment where the mean scores for enjoyment recorded for the LOW and MID complexity VW tasks were much higher than the FTF equivalents. However, the same was not true of the HIGH complexity tasks where mean scores were more closely matched (Table 31 above).

4.1.2.1 Task complexity effects on learner perceptions

A one-way repeated measures ANOVA was run on the data to investigate the effect of task complexity on participants preference for completing tasks in the VW or FTF for each measure. There was one outlier found for the *task difficulty* measure at the HIGH level, which had a studentized residual value of 3.88. Investigating the outlier in further detailed revealed that it related to a participant who did not differentiate between the perceived difficulty of the HIGH complexity tasks. That is, for this measure, 18 of the 20 participants responded with a value of 1, indicating that the VW task was more difficult than the FTF equivalent, one participant responded with a value of 2, and the outlier responded with a value of 3. The data was recorded correctly and was therefore not removed.

The data were not normally distributed for any of the measures, as assessed by Shapiro-Wilk's test of normality on the studentized residuals. Visual inspection of Q-Q plots, however, revealed that the data was reasonably normally distributed. The data was therefore not transformed. Mauchly's test of sphericity revealed that the sphericity was assumed for all measures (Table 34).

Table 34: Mauchly's test of sphericity results for all task comparison questionnaire measures.

Measure	Mauchly's W	Approx. Chi-Square	df	Sig.
Pushed output	.81	3.75	2	.15
Enjoyable	.97	.48	2	.79
Difficult	.94	1.03	2	.60
Learning	.99	.20	2	.90

There was a statistically significant main effect for task complexity on two of the four measures: task difficulty, $F(2, 38) = 25.28$, $p < .001$, and the learning potential of the tasks, $F(2, 38) = 12.00$, $p < .001$. There was no statistically significant main effect of task complexity for the *pushed output*, $F(2, 38) = 0.39$, $p = .68$ and *enjoyment* measures, $F(2, 38) = 1.69$, $p = .20$. This indicates that for these two measures, task complexity had no significant effect on participants' preference for completing the task FTF or within the VW. All tasks were perceived as equally demanding in terms of output requirements, and the VW tasks were perceived as more enjoyable than the FTF tasks regardless of task complexity (lower mean scores indicating a preference of the VW modality).

Inspection of pairwise comparisons for task difficulty mean scores revealed that the HIGH complexity tasks' mean scores were statistically significantly lower than both the LOW complexity tasks (mean difference = 1.25 points, $p < .001$) and the MID complexity tasks (mean difference = 1.9, $p < .001$). This suggests a bias towards perceiving the VW HIGH task as more difficult than the FTF equivalent. A result which supports findings related to the post-task questionnaire measure, where it was found that VW HIGH complexity task was perceived as particularly cognitively demanding (section 4.1.1.2).

Pairwise comparisons for the learning potential measure revealed similar results to those found for task difficulty. A statistically significant effect of task complexity on participants' perception of the learning potential of tasks was found between the LOW-HIGH, ($p = .01$) and MID-HIGH, ($p < .001$) levels. The mean score recorded for the HIGH complexity task was 0.95 points lower than the LOW mean score and 1.25 points lower than the MID mean score. This indicates that the participants perceived the VW HIGH task to

hold a particularly high learning potential over the FTF equivalent.

4.1.3 Open-ended question responses

The previous section presented results for the Likert-like questions of the individual post-task and task-comparison questionnaires. Subsequently, this section provides results from the open-ended questions on both questionnaires.

4.1.3.1 Post-task questionnaire responses

For the post-task questionnaire, three open-ended questions were employed. The first asked learners to consider what they learnt from undertaking the task. The second question asked learners to make comments on any positive elements of the tasks. The final question asked learners to consider what elements of the tasks they would improve, which was interpreted in two ways, 1) regarding their own performance, and 2) regarding ways in which the task could be designed to promote language development more efficiently.

In the following sections, responses are explored in relation to a specific coding schema that was created based on patterns found in the open-ended responses. Codes fall under the superordinate categories of 1) task-related comments, 2) language-related comments and 3) environment-related comments. Findings are presented in order, exposing any differences in responses for the FTF and VW tasks. As a review, the following table presents the codes used in this study (

Table 35). The total number of responses for each code can also be seen in Appendix 2 (individual post-task questionnaires) and 3 (post-task comparison questionnaires).

Table 35: Coding scheme employed for comments to the open-ended questions of the individual post-task questionnaire.

Task For any comments related to the task itself.	Language (code complexity) For any comments related to learning gains or language-specific comments.	Environment For any comments that mention the environment itself (i.e. FTF or VW affordances.)
fun	authentic	fun
simple	simple	easier
difficult	difficult	WTC
	communication	technical
	pronunciation	affordances
	words	
	grammar	
	phrases	
	translation	
	listening	
	non-verbal	
	JP	
	ZPD	

In terms of the total possible number of responses, the 20 participants completed three FTF and three VW tasks, and there were three open-ended questions on each questionnaire, this means that for each mode, there was a possible total of 180 responses which would need to be coded (3 tasks x 3 questions x 20 participants). However, there was a total of 135 responses recorded for the FTF tasks, and 127 for the VW tasks.

Codes were not fixed to a single open-ended question but applied to any response where their inclusion was deemed appropriate. Thus it was possible that certain responses would be coded with codes from all three superordinate groups and even multiple subordinate categories of the same group. In the following section, percentage values refer to the per cent of total responses for each mode. For example, if there were 12 responses coded *task-fun* in response to open-ended questions for the FTF tasks, this is 8.89% of the total recorded

responses for that modality (12 / 135 = 0.09).

4.1.3.1.1 Task-related codes

The task code "task-fun" corresponds somewhat with the task "enjoyment" measure above. The result of the post-task questionnaire measure for task enjoyment was that tasks were comparable in terms of their enjoyment regardless of modality. A similar result is found here: There are an almost equal number of responses that were tagged with the code "task-fun," (FTF n = 12, VW n = 13) (Table 36), therefore, reflecting the results found above: Tasks were considered enjoyable regardless of modality.

Table 36: Task-related codes applied to open-ended question responses for FTF and VW tasks.

Code	FTF	VW
Fun	12 (8.89%)	13 (10.24%)
Simple	28 (20.74%)	7 (5.51%)
Difficult	4 (2.96%)	10 (7.87%)

Note. Percentages relate to the ratio of each code to the total number of responses for the superordinate group.

Example responses are provided in Table 37 which highlight a trend in results. For the FTF tasks, responses that were coded with task-fun were also often also coded with task-simple, whereas for the VW tasks, task-fun was seen more in combination with responses regarding the affordances of the environment. In other words, it seems that "fun" or enjoyment of a task was not determined by any particular task condition. Instead, there are affective, cognitive, and mode-based influences on participants' task enjoyment perceptions.

Table 37: Examples of responses coded as "Task-fun."

Participant Number	Task	Comment
5	VW MID	こういう活動を通して学ぶことが話す練習するより楽しかった。[It was fun to learn this way (through completing an activity) rather than by assigning us to speak English for speaking practice.]
1	FTF MID	簡単で楽しかったです。[It was easy and fun.]
4	FTF LOW	英語が苦手なのにこの活動でどうすればいいかがわかりやすくて目標達成ができたので楽しかった。[It was fun for me because even though I am not strong at English, I could understand what to do in this activity and I was able to achieve the task goals.]
6	VW LOW	ゲームを通して英語が学べることが素晴らしいと思った。 [I think it is amazing that I can learn English by playing a game.]

Findings suggest that, as predicted, the FTF tasks were considered simpler than their VW counterparts. A total of 28 responses in the individual post-task questionnaire open-ended questions were coded with the "task-simple," which is 20.74% of all responses recorded for the FTF tasks. Compared to this, for the FTF tasks, only 4 responses (2.96%) were coded with task-difficult. Conversely, for the VW tasks, although more responses were coded with "task-difficult" than "task-simple," there is less of a clear divide. Task-difficult was applied to 10 responses (7.87%), and task-simple was applied to only 7 (5.51%). This suggests that the difficulty of VW tasks was not necessarily perceived as being high for all participants.

Additionally, 8 of the 10 *task-difficult* codes were applied to responses for the VW HIGH task, and two for the VW MID task, which suggests there was a clear divide regarding which task participants considered difficult. This finding corresponds with the closed-questions of the questionnaire which suggested that of all the tasks the participants completed, the VW HIGH task was the most cognitively demanding, and required the most mental effort to complete.

4.1.3.1.2 Language-related codes

For language-related codes, there were a number of differences found between those codes applied to FTF and VW task responses. An overview is provided in Table 38.

Table 38: Language-related codes applied to open-ended question responses for FTF and VW tasks.

Code	FTF		VW	
Authentic	5	(3.70%)	10	(7.87%)
Simple	14	(10.37%)	14	(11.02%)
Difficult	12	(8.89%)	29	(22.83%)
Communication	27	(20.00%)	38	(28.35%)
Pronunciation	2	(1.48%)	1	(0.79%)
Words	20	(14.81%)	11	(8.66%)
Grammar	32	(23.70%)	8	(6.30%)
Phrases	4	(2.96%)	6	(4.72%)
Translation	0	-	1	0.79%
Listening	4	2.96%	0	-
Non-verbal	6	4.44%	0	-
JP	0	-	3	1.57%
ZPD	0	-	3	2.36%

Note. Percentages relate to the ratio of each code to the total number of responses for the superordinate group.

There were twice as many responses coded with *language-authentic* for the VW tasks than the FTF tasks which suggests that the perceived authenticity of communication in the VW domain was higher than for the FTF tasks. Of note here is that 6 of the 10 VW task responses that were coded as *language-authentic* appeared in response to the VW MID task (directions task), which suggests that traversing the immersive environment may have been a substantial contributing factor. Participant 13's response is an illustrative example of this:

実際に道案内しているような感じだった。

[*It felt like I was actually giving directions in real life*.]

Codes for language simplicity or difficulty appear different to the task-related codes

explored above. For the FTF tasks, there is an almost equal number of responses which mention the simplicity (n = 14) or difficulty (n = 12) of language encountered in the tasks, however for the VW tasks, there is a sharp increase in the number of responses that were coded as *language-difficult* (n = 29, 22.83%). Again, the majority of responses coded with language-difficult appeared in the VW HIGH post-task questionnaire (n = 18).

One interesting finding is that participants perceived the two modes of communication to promote different kinds of language use. For the FTF tasks, a number of responses were coded as *language-grammar*, *-words*, and *-phrases*, suggesting that the FTF tasks were mostly perceived as being suitable for promoting grammar-oriented skills. It seems that the simplicity of the tasks or participants' familiarity with the language used in the tasks may have attributed to this trend. An example of how task simplicity effected this result can be seen in Participant 5's response to the FTF LOW task:

単語だけでも英語でコミュニケーションができると思った。

[*I felt like I can communicate in English by using words only.*]

The response suggesting that the task was simple enough to be completed through the use of words only, requiring only surface-level interaction, and possibly no deeper communication other than describing a character. Additionally, the same participant's response to the FTF MID task suggests that the task was so simple that they only learnt a single word by completing the activity:

"Alley と言う単語を学んだ。最初は valley のスペルミスだと思った。"

[*I learnt the word 'alley.' I thought it was a misspelling of 'valley' at first."*]

In comparison, for the VW tasks, responses mentioned that communication was fostered more, where 38 responses (28.35%) were coded with *language-communication*. The VW tasks are therefore considered more useful in promoting communication between participants, with less of a focus on grammar or "language practice." This result may indicate why there is a higher number of responses coded with *language-authentic* for the VW tasks.

Additionally, and connected to the above finding is the result for the code *language-ZPD.* There were only three responses in total coded this way, but all three were found in response to VW tasks. These are:

Participant 10, VW LOW: 相手が言った事を使おうとした。 [*I tried to use the words my partner said.*]

Participant 19, VW LOW: パートナーが助けてくれた時に言い回しがわかった。質問を考えるのが良かった。[*I learned how to speak when my partner helped me. It was good to think about questions.*]

Participant 2, VW HIGH: 最初は説明できなかったけど、パートナーが話していた時にどう説明すればいいかわかった。 [*I didn't know how to explain something, but I learnt how to say it during my partner's turn*.]

Regarding the language learning benefits of the two modes of communication then, results suggest that the FTF tasks promoted learners to engage in "language practice" activities, where learning was typically seen to focus on grammar, vocabulary, or other lexical items. On the other hand, the VW tasks seemed to promote communicative skills, interaction, and cooperation between participants, thus more in line with the philosophy of a CLT approach to language development.

Finally, worth highlighting here is the difference in results for the code *language-non-verbal.* Responses coded this way were only found in FTF post-task questionnaires, which may confirm that the affordances of the FTF mode allowed learners to rely on body language more than in the VW. Such may be interpreted as the virtual domain proving to be a hindrance on natural communication, or, alternatively, that the VW forced participants to rely on verbal interaction instead. Both of these views are suggested in participant responses. Participant 20 on the FTF MID task:

パートナーと実際に話した方が説明しやすいと思う。 [*I think it is easier to explain something to my partner when I am face to face with them*].

And the opposite interpretation by Participant 8 in his response to the VW HIGH task (explored in more detail in Section):

ゲームで行う方は、ボディーランゲージが使えない分、純粋に英語力に頼ることになり、勉強になった。[*When doing the online version, we cannot read our partner's body language, so we have to rely on our English skills. This means that we learn more when doing*

the virtual world version.]

4.1.3.1.3 Environment-related codes

Among similarities in terms of the number of environment-related responses, one sub-category code "*environment-technical*" was applied overwhelmingly to responses that mentioned the VW (Table 39). 12 responses for the FTF tasks mentioned technical issues, whereas there were 27 responses for the VW tasks. Of these, the most common technical issues for the FTF tasks were found with the LOW complexity task (n = 6) due to a lack of clarity in the picture provided on the worksheet:

Participant 3: 絵の中でキャラクターは何しているか、いくつかわかりにくいのがあった。 [*It was hard to tell what the characters were doing in some of the pictures.*]

However, note that the FTF LOW task was considered the least mentally demanding of all tasks. Thus, the technical problems associated with this task did not appear to affect task difficulty perceptions. This may be because the task was simple enough (in terms of language and interactional requirements) that the technical issues were not considered a problem.

Table 39: Environment-related responses for FTF and VW tasks.

Code	FTF		VW	
Easier	5	(3.70%)	2	(1.57%)
WTC	2	(1.48%)	2	(1.57%)
Technical	12	(8.89%)	27	(21.26%)
Affordances	12	(8.89%)	12	(9.45%)
Fun	3	(2.22%)	9	(7.09%)

Note. Percentages relate to the ratio of each code to the total number of responses for the superordinate group.

For the VW tasks, the HIGH complexity task appeared to cause the most technical difficulties with a total of 16 responses coded as such for this task. Looking qualitatively at responses reveals that controlling avatars and manipulating the world in response to their partner's instructions were prime sources of confusion and technical difficulties for this particular task:

Participant 6: 操作方法は学んでなかったので難しかった。 [*It was difficult because we hadn't been taught how to control the characters in the game.*]

Participant 17: アイテムを逆さまに置くのが難しかった。 [*It was hard to place some of the items that were upside down.*]

4.1.3.1.4 Summary of open-ended question responses for the individual post-task questionnaires.

In summary to this section, responses to the open-ended questions of the individual post-task questionnaires suggest that there were a number of differences between how participants perceived the tasks in terms of three categories: task-related, language-related, and environment-related responses.

For codes applied to responses that were task-related, there was a clear trend for participants to mention that the FTF tasks were simpler, and presumably easier to complete than VW equivalents. For the VW tasks, although more task-difficult codes were recorded, the ratio of simple and difficult responses was much lower than for the FTF tasks. Language-related codes revealed that the FTF tasks seemed to promote skills for the acquisition of specific language items such as grammatical concepts or vocabulary. For the VW tasks, language-related codes were clustered more around the concepts of authentic language use and communication as a specific skill. Finally, for the environment-related codes, there were a large number of responses that mentioned specific technical difficulties that the VW tasks presented. Reasons for this and the affective and cognitive implications of this is explored in the Discussion section (Section 5.3).

4.1.3.2 Task-comparison questionnaire responses

A single open-ended question was given to participants on the post-task comparison questionnaire. It was designed to provide participants with a space to provide further explanation or comments regarding their choices, or general comments regarding the task-pair that they just completed. This section explores recorded responses through a similar coding system that was employed for the individual post-task questionnaires.

In terms of the total possible number of responses, the 20 participants completed three post-task comparison questionnaires and so there was a potential for 60 responses to be made (20 x 3 = 60). However, the actually recorded volume of responses was only 43. In the

following section, percentage values refer to the per cent of total responses over all three post-task comparison questionnaires. For example, if there were 8 responses coded *FTF-cognitive-low* this is 18.60% of the total recorded responses (8 / 43 = 0.19). The total number of codes can be seen in Table 40.

Table 40: Number and percentage of codes used in the post-task comparison questionnaires.

Code	FTF		VW	
cognitive-low	8	(18.60%)	2	(4.65%)
cognitive-high	0	-	15	(34.88%)
WTC	0	-	3	(6.98%)
technical	1	(2.33%)	10	(23.26%)
affordances	3	(6.98%)	26	(60.47%)
positive	2	(4.65%)	25	(58.14%)

Note. Percentages relate to the ratio of each code to the total number of responses for the superordinate group.

There is a large discrepancy in the number of responses for each code based on modality. Of 43 responses, only 14 FTF codes were recorded, compared to a total of 81 codes for the VW tasks. This suggests that the VW tasks were perhaps more memorable, or at least worth additional comments from the participants compared to the FTF tasks. Looking at the codes in more detail reveals that, as discussed already, responses generally suggested that the FTF tasks were less cognitively demanding than the VW tasks (8 FTF-cognitive-low codes and 15 VW-cognitive-high codes).

Three responses were coded with WTC (an abbreviation of Willingness to Communicate), which was used when the responses seemed to suggest that a particular task motivated learners to communicate. It is worth noting here that all responses coded this way fall under the VW superordinate categorisation, and additionally, on the LOW complexity task-comparison questionnaire. One example of this was simply "オンラインでやる方は緊張しなかった。 [*I wasn't so nervous when I did the VW version*]" (Participant 9), suggesting that the addition of a tool for mediating conversation (in this study: computers, mics and headsets), may have been a positive influence on reducing social anxiety, leading

to increased WTC. In terms of RQ2, and the affective effects of communicating in a VW, this point has significant importance. Results so far have shown that the VW tasks were considered more enjoyable than the FTF tasks, and here there is possible evidence that the anonymizing "avatar-effect" may have been a contributing factor.

The affordances of the VW tasks was mentioned in 26 (60.47%) of the responses, which connects to the "positive" code, appearing in 25 responses (58%). The affordances of the VW were thus considered an additional influential component in determining task enjoyment. Examples suggest that this enjoyment of the VW tasks came from the ability to explore the terrain or agency that the VW afforded, such as in Participant 18's comment on the LOW complexity task-comparison questionnaire: "マイクラの中でキャラクターを探すのが楽しかった。[*It was fun to search for the characters in Minecraft.*]" Additionally, the perceived high level of mental effort required to complete the VW HIGH task was not considered a drawback to completing this task, but alternatively a positive element: **Participant 1, HIGH complexity task-comparison questionnaire:** オンライン版で活動した時はもっと英語で話す機会があった。難しかったけど楽しかった。[*We had more opportunities to use English with the online version. It was difficult but fun*.]

Participant 15, MID complexity task-comparison questionnaire: マイクラでキャラクターを操作するのが難しかったけど環境がとてもリアルだったので紙でやるよりは楽しかった。 [*It was difficult to control the character in Minecraft, but the environment was very immersive so much more fun than doing the paper version*.]

4.1.4 Summary of participants' task difficulty perceptions

Having reviewed the results of the post-task questionnaires, I now return to the topic of perceived task difficulty. That is, separate from the researcher's prediction of task complexity, task difficulty is a subjective perception of how difficult a task is based on a number of subjective, individual factors. In this study, and in keeping with the measurement of task difficult found in the literature (Kim, 2009; Révész, 2011; Robinson, 2007; Sasayama, 2013), a Cognitive Load Subjective Experience Questionnaire was utilised. It is possible to infer perceived task difficulty by comparing the results of the post-task and post-task comparison questionnaires.

There is a strong positive correlation between perceived mental effort and task difficulty, as recorded on the post-task questionnaire. Learners perceived themselves as exerting more mental effort on the tasks that they also perceived to be more difficult. These results suggest that the LOW complexity tasks posed the least amount of cognitive demand, and the HIGH complexity tasks were the most complex and cognitively demanding. There was also a two-way interaction between modality and task complexity for the task difficulty measure, which indicated that the VW HIGH complexity task was, as predicted, the most cognitively demanding of all six tasks. Overall then, the LOW tasks were perceived to be the simplest task by all measures of cognitive load, and the HIGH tasks were perceived to be the most demanding.

Responses to the Likert-like questions on the post-task comparison questionnaire corroborated with the post-task questionnaires. There was a statistically significant effect of task complexity on participants' perception of the learning potential and difficulty of the tasks, where the VW HIGH task was considered to be particularly cognitively demanding, yet at the same time, hold a greater learning potential than the FTF equivalent at this level. This finding, along with results of the coded open-ended questions data, lends support to the notion that the VW tasks had a positive influence on participants' motivation. On the post-task comparison questionnaire mean scores for *task enjoyment* were all lower than the median value of 3, which again indicates that despite being more cognitively demanding, the VW tasks were considered more enjoyable than the FTF tasks for all levels of task complexity.

Findings, therefore, lend support to the claims of Robinson's CH in that operationalization and manipulation of task conditions was effective in altering a task's complexity. This was verified via inspection of quantitative data which suggested that the participants' perception of task difficulty matched closely with predefined task complexity assumptions. Knowing that task complexity was perceived as predicted, it is now possible to explore quantitative speech data and attempt to answer RQ1.

4.2 EFFECT OF MODALITY AND TASK COMPLEXITY ON LEARNER TASK PERFORMANCE

The following section explores how modality and task complexity affected learners' task performance, thus answering RQ1. The analysis was conducted systematically, according to the different CAF measures. Thus, in the following sections, complexity

measures are focused on first, then accuracy measures, and finally fluency measures. After presenting results for these three dimensions of performance, a final summary section highlighting significant findings is presented.

4.2.1 Complexity measures

4.2.1.1 Syllables per utterance

Mean scores for each participant over the six tasks are displayed in Table 41 and graphically in Figure 21. The highest mean score was recorded for the LOW complexity tasks (FTF - 4.95 syllables per utterances, SD 1.08; VW - 4.03 syllables per utterance, SD = .93). The lowest mean scores were recorded for the HIGH complexity tasks for both modes of communication. The results of descriptive statistics for this measure reveal that in general, the mean scores for the FTF tasks were higher than the VW tasks for all levels of task complexity. The next stage of analysis was to run a two-way repeated measures ANOVA on the data.

Table 41: Mean scores and SD for syllables per utterance measure.

Task	Mean	Std. Deviation
FTF LOW	4.95	1.08
FTF MID	3.98	0.72
FTF HIGH	3.28	0.74
VW LOW	4.03	0.93
VW MID	3.58	0.90
VW HIGH	3.09	1.16

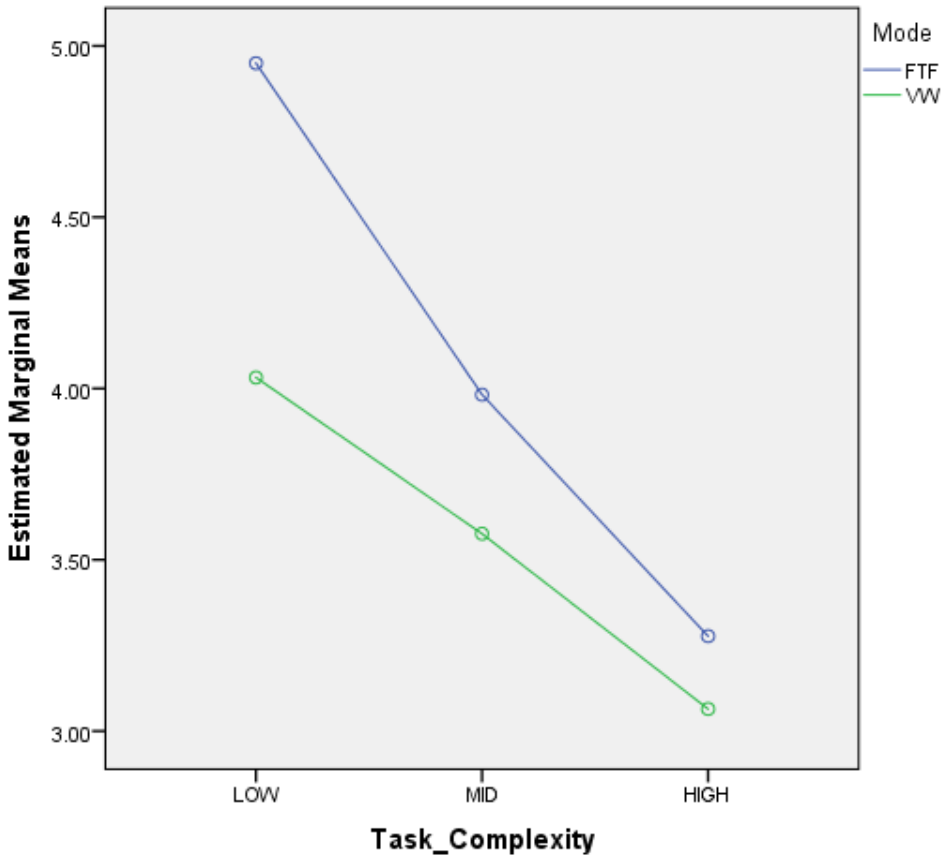


Figure 21: Estimated marginal means of the syllables per utterances measure for all six tasks

A two-way repeated measures ANOVA was run to determine whether there was an interaction between modality and task complexity on the number of syllables participants produced per utterance. There were no outliers, as assessed by examination of studentized residuals for values greater than ±3. Syllables per utterances scores were normally distributed for all tasks, as assessed by Shapiro-Wilk's test of normality on the studentized residuals ($p > .05$). Mauchly's test of sphericity indicated that the assumption of sphericity was met for a two-way interaction between modality and task complexity for this measure, χ^2 (2) = 2.73, $p = .26$. Visual inspection of profile plots in Figure 21 revealed that there was possibly an exponential interaction between modality and task complexity. Inspection of the results of the two-way repeated measures ANOVA revealed that there was a statistically significant two-way interaction, $F(2, 38) = 4.0$, $p < .05$. However, partial Eta squared results suggested that the amount of variance on output complexity can be attributed to modality and task complexity main effects more than the interaction of both, particularly task complexity which had an effect size of 0.65 (see Table 42).

Table 42: Two-way repeated measures ANOVA for modality and task complexity on words per utterance *complexity measure.*

Source	Sum of squares	df	Mean square	F	p	Partial Eta sq.
Mode	7.86	1	7.86	19.28	.000	0.50
Task Complexity	34.92	2	17.46	35.79	.000	0.65
Mode x Task Complexity	2.65	2	1.33	4.00	.03	0.17

Examining the main effect for modality via pairwise comparisons suggested that the mean difference in the number of syllables per utterance was 0.51 higher in the FTF mode over the VW mode, 95% CI [.48 to 1.36], $p < 0.001$. For task complexity, the main effect revealed that the LOW complexity task allowed participants to produce statistically significantly more syllables per utterance than both the MID and HIGH complexity tasks (Table 43). Additionally, there was a statistically significant difference in mean scores between the MID and HIGH complexity tasks where learners produced 0.61 syllables per utterance more when completing the MID complexity task, 95% CI [.21 to 1.00], $p = 0.002$. However, as the two-way repeated measures ANOVA revealed an interaction between the two factors, simple main effects were explored in more detail.

Table 43: Pairwise comparisons for task complexity main effect on syllables per utterance.

Task Complexity	Task Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower Bound	Upper Bound
LOW	MID	0.71*	.18	.002	.25	1.17
	HIGH	1.32*	.14	.000	.95	1.69
MID	HIGH	0.61*	.15	.002	.21	1.00

Notes. The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

4.2.1.1.1 Simple main effects of modality on syllables per utterance

To investigate the simple main effects of modality, three comparisons of the data were conducted at the LOW, MID and HIGH levels of task complexity, the results of which

are available in Table 44. Results revealed a significant simple main effect for modality at the LOW and MID complexity levels after applying a Bonferroni adjustment to the level of statistical significance ($p < .05$). However, the same was not true for the HIGH task complexity ($p = .30$).

Table 44: Pairwise comparisons for simple main effects of modality on syllables per utterance

Task Complexity	(I) Mode	(J) Mode	Mean Difference (I-J)	Std. Error	Sig[a]	95% CI[a]	
						Lower Bound	Upper Bound
Low	FTF	VW	.92*	.21	.00	.48	1.36
Mid	FTF	VW	.41*	.15	.02	.09	.72
High	FTF	VW	.21	.20	.30	-.21	.63

Notes. The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

Interpreting these results, it appears that as task complexity increased, participants struggled to produce complex output. This result appears contrary to the CH, in that as task complexity increases, learners' output complexity are predicted to increase accordingly to meet the requirements of the task. The results above suggest that with lower task complexities, participants output became more statistically significantly more complex, particularly when conducting the LOW complexity task. In terms of the effect of modality, the increase in output complexity with LOW complexity tasks is only seen with the FTF tasks. For the VW tasks participant output complexity appears to be hindered, even with the LOW complexity task.

In the following section, the same measure is explored in terms of simple main effects of task complexity.

4.2.1.1.2 Simple main effects of task complexity on syllables per utterance

For the FTF tasks, the assumption of sphericity was met, $\chi^2(2) = 5.06$, $p = .08$. Results suggest that there was a statistically significant effect of task complexity on participants' task performance in terms of syllables per utterances when completing the FTF tasks, $F(2, 38) = 32.92$, p = < .001. Subsequently, a significant difference in mean scores was found between

all levels of complexity. There was an overall decrease in mean syllables per utterance by 1.67 syllables between the LOW and HIGH complexity tasks, 95% CI [1.05, 2.3], $p < .001$.

Table 45: Pairwise comparisons for the simple main effect of task complexity on syllables per utterance for FTF modality.

(I) Complexity	(J) Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower	Upper
LOW	MID	.97*	.25	.003	.25	1.69
	HIGH	1.67*	.21	.000	1.05	2.30
MID	LOW	-.97*	.25	.003	-1.69	-.25
	HIGH	.71*	.15	.001	.26	1.16

*The mean difference is significant at the .05 level.
[a]Adjustment for multiple comparisons: Bonferroni.

For the VW tasks, sphericity was also met, $\chi^2(2) = .24$, $p = .89$ and a significant effect of task complexity on syllables per utterance was found, $F(2, 38) = 11.97$, $p < .001$. Examination of pairwise comparisons revealed that the only statistically significant difference in means was found between the LOW ($M = 4.03$, $SD = .93$) and HIGH ($M = 3.06$, $SD = 1.16$) complexity tasks where there was a decrease in mean syllables per utterance by .97 syllables, 95% CI [-1.46, -.48], $p < .001$. Additionally, the difference in mean scores between the MID and HIGH complexity tasks also approached statistical significance ($p = .07$).

Table 46: Pairwise comparisons for the simple main effect of task complexity on syllables per utterance for VW modality.

Complexity	**Complexity**	**Mean Difference**	**Std. Error**	**Sig[a]**	**95% CI[a]**	
					Lower	Upper
LOW	MID	.46	.20	.10	-.07	.98
	HIGH	.97*	.19	.00	.48	1.46
MID	HIGH	.51	.21	.07	-.03	1.05

*The mean difference is significant at the .05 level.
[a]Adjustment for multiple comparisons: Bonferroni.

In summary, then, results suggest that as task complexity increased, learners produced less complex output with fewer syllables per utterance. This was true regardless of modality. Task complexity possibly had a negative effect on oral complexity, a finding which is opposite to the claims of the Cognition Hypothesis, but that lends support to the trade-off effect of the LACM in that as task complexity increases, participants were unable to attend to their linguistic output.

4.2.1.2 Number of different words

Mean scores for each participant over the six tasks are displayed in Table 47, and a visual representation of mean scores is available in Figure 22. The results of descriptive statistics for this measure reveal that in general, the mean scores for the VW tasks were higher than the FTF tasks for all levels of task complexity. The highest mean score was recorded for the HIGH complexity room decoration task. However, the standard deviation for this task is also large, indicating that this task produced the most varied performance amongst participants (VW HIGH mean = 99.8, SD = 53.48). Conversely, the lowest mean score was recorded for the FTF HIGH complexity task.

Table 47: Mean scores and SD for different words measure.

	Mean	Std. Deviation
FTF LOW	59.48	20.13
FTF MID	67.45	27.43
FTF HIGH	58.13	28.87
VW LOW	65.25	20.21
VW MID	71	30.61
VW HIGH	99.8	53.48

A two-way repeated measures ANOVA was run to determine whether there was an interaction between modality and task complexity on the number of different words that participants used. The data was not normally distributed for three of the six tasks, as assessed by Shapiro-Wilk's test of normality on the studentized residuals. Additionally, two outliers were recorded, which had studentized residual values of 3.05, recorded for the FTF version of the directions task and 3.19, recorded for the FTF version of the room decoration task. The

outliers were recorded for the same participant, who had a higher than average proficiency level. Outliers were removed, and statistical tests ran twice. However, statistical significance was not affected, and so the data with outliers included was used for further investigation.

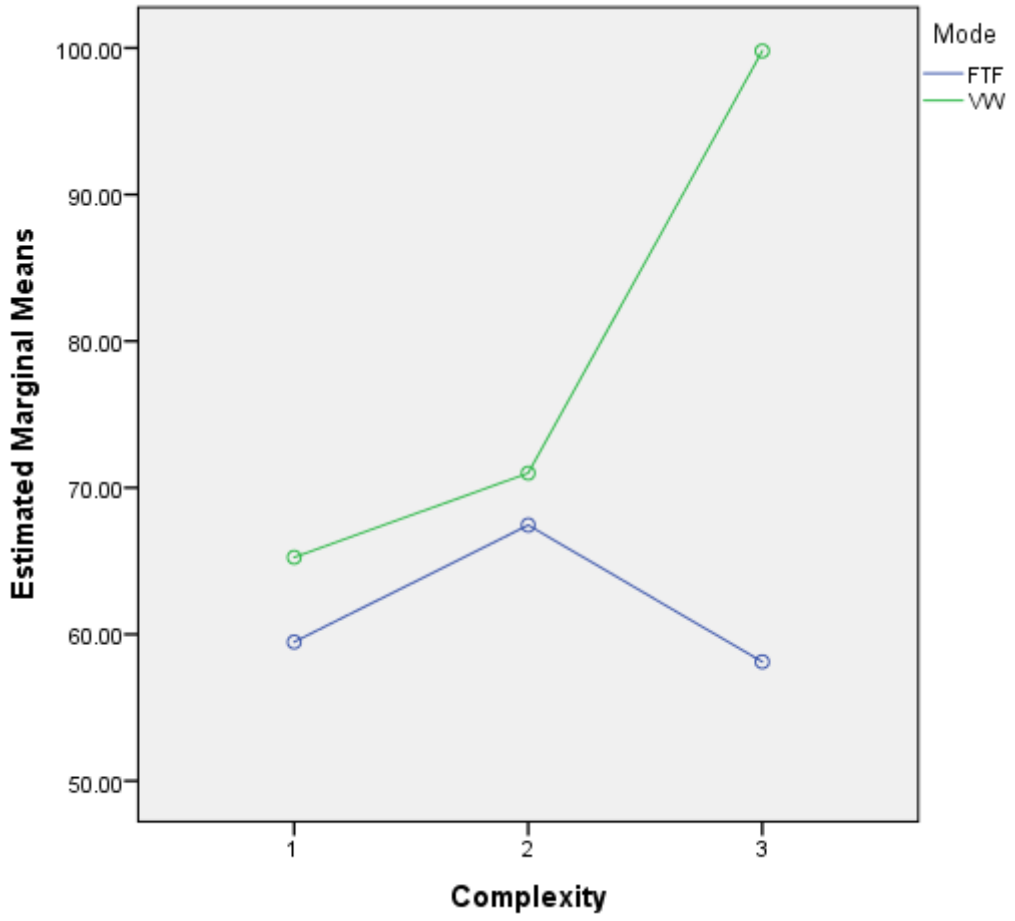


Figure 22: Estimated marginal means for the different words measure.

Mauchly's test of sphericity indicated that the assumption of sphericity was violated for the two-way interaction, χ^2 (2) = 6.64, p = .04. Epsilon (ε) was .76, as calculated according to Greenhouse-Geisser and was used to correct the two-way repeated measures ANOVA. Visual inspection of profile plots in Figure 22 revealed that the two lines generated for modality were approximately parallel for the LOW and MID complexity tasks. However, a large discrepancy was recorded between mean scores for the two HIGH complexity tasks, where results diverge. The FTF room decoration promoting participants to use the least amount of different words of all the three FTF tasks, whilst the VW equivalent pushed participants to output the largest volume of different words (mean = 99.8). Results of the

two-way repeated measures ANOVA revealed that there was a statistically significant two-way interaction, F (1.53, 29.04) = 17.16, $p <$ .05 (Greenhouse-Geiger corrected). Additionally, in terms of main effects, there was statistical significance found for both modality and task complexity.

In terms of the main effect, pairwise comparisons for modality suggested that the mean difference in the number of different words used by participants was 17 words higher in the VW mode over the FTF mode, 95% CI [10.11 to 23.89], $p <$ .001. For task complexity, although there was a statistically significant main effect, pairwise comparisons between the three different levels of complexity did not reveal any statistical significance in mean scores (Table 48).

Table 48: Pairwise comparisons for the main effect of modality on the different words measure

(I) Complexity	(J) Complexity	Mean Difference (I-J)	Std. Error	Sig[a]	95% CI[a]	
					Lower Bound	Upper Bound
LOW	MID	-6.86	3.92	.29	-17.15	3.43
	HIGH	-16.60	6.60	.06	-33.91	.71
MID	LOW	6.86	3.92	.29	-3.43	17.15
	HIGH	-9.74	5.29	.24	-23.61	4.14

Notes. *The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

4.2.1.2.1 Simple main effects of modality on the number of different words produced

Three comparisons of the data were conducted at the LOW, MID and HIGH levels of task complexity, results of which are available in Table 49. The analysis revealed no statistically significant simple main effect for modality at the LOW and MID complexity levels. However, there was a statistically significant simple main effect for the HIGH complexity task where the mean number of different words spoken participants completing the VW task was on average 41 (95% CI, 25.85 to 57.50) words higher than the FTF task, F(1, 19) = 30.37, $p <$ .05).

Table 49: Pairwise comparisons for simple main effects of modality on the number of different words participants produced

Task Complexity	(I) Mode	(J) Mode	Mean Difference (I-J)	Std. Error	Sig[a]	95% CI[a]	
						Lower Bound	Upper Bound
Low	FTF	VW	-5.78	3.39	.11	-12.88	1.33
Mid	FTF	VW	-3.55	4.13	.40	-12.20	5.10
High	FTF	VW	-41.68	7.56	.00*	-57.50	-25.85

Notes. *The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

Based on the claims of the CH, as task complexity increases, there should be an increase in learner output complexity as they are pushed to higher levels of linguistic complexity. This seems to be the case for the LOW and MID complexity tasks where there is a trend for learners to output a larger number of different words in the MID complexity tasks compared to the LOW complexity tasks. Subsequently, for the HIGH tasks, this trend only continues with the VW tasks, as the VW HIGH task pushes learners to produce the highest volume of different words, matching the predictions of the CH. There is however an issue with the FTF HIGH task in that there is a drop in output complexity compared to the other FTF tasks, a phenomenon contrary to the predictions of the CH. Rather than modality, then, the cause of this drop could be related to task conditions. Further analysis is done in the discussion chapter below.

4.2.1.2.2 Simple main effects of task complexity on the number of different words participants produced

For the FTF tasks, the assumption of sphericity was met, $\chi^2(2) = 3.91$, $p = .14$. Results suggest that there was a statistically significant simple main effect of task complexity on participants' task performance regarding the different number of words they produced, $F(2, 38) = 3.78$, $p = .03$. However, a significant difference in mean scores was only found between the MID and HIGH complexity tasks. There was an overall decrease in the number of different words participants produced by 9.33 words between the MID and HIGH complexity tasks for this mode, 95% CI [2.02, 16.63], $p = .01$ (See Table 50).

Table 50: Pairwise comparisons for the simple main effect of task complexity on the number of different words for FTF modality.

(I) Complexity	(J) Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower	Upper
LOW	MID	-7.98	3.85	.16	-18.09	2.14
	HIGH	1.35	4.21	1.00	-9.70	12.40
MID	HIGH	9.33*	2.78	.01	2.02	16.63

Notes. The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

For the VW tasks, sphericity was violated, χ^2 (2) = 9.50, p = .01 and a significant effect of task complexity on the number of different words was found, F(2, 38) = 9.64, p < .001. Examination of pairwise comparisons revealed that statistically significant differences between the mean scores for both LOW and HIGH (mean difference of 34.55 words) and MID and HIGH (mean difference of 28.8 words) complexity tasks where the HIGH complexity task pushed learners to produce the highest number of different words (Table 51).

Table 51: Pairwise comparisons for the simple main effect of task complexity on the number of different words for VW modality.

Complexity	Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower	Upper
LOW	MID	-5.75	5.55	.94	-20.33	8.83
	HIGH	-34.55*	10.47	.01	-62.03	-7.07
MID	HIGH	-28.80*	8.53	.01	-51.19	-6.41

Notes. The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

In summary, for this measure results suggest that the mean number of different words that participants produced were significantly different when completing tasks in the FTF and VW modalities. However, for the FTF tasks, a significant difference was only found between the MID and HIGH complexity tasks. The MID complexity task promoted learners to produce the highest number of words out of the three tasks for this modality, which was unexpected. Based on the claims of the CH, the HIGH complexity was expected to push

learners to the greatest level of output complexity, but conversely, for the FTF mode, this task promoted the worst performance, with learners only producing 58.13 different words when carrying out this task. In general, it seems that tasks completed in the FTF mode did not push learners to produce a high number of different words.

For the VW tasks, significance was found between LOW and HIGH and MID and HIGH complexity tasks, but as with the FTF tasks, the mean scores for the MID and LOW complexity were not significantly different. The HIGH complexity task pushed learners to produce the highest number of different words on average, where the highest overall mean score was recorded (99.8 words). This result is in keeping with expectations; that with increased task complexity, learners were pushed to produce more complex output. It also corroborates with the findings presented above on participants' perception of vocabulary difficulty. That is, the vocabulary encountered in the HIGH complexity tasks was perceived to be statistically significantly more difficult than the MID and LOW complexity tasks.

One issue is that the interaction between modality and task complexity here is opposed to the results for the syllables-per-utterance measure of output complexity. One reason for this may be that whilst the interaction of modality and task complexity negatively affected *structural* complexity (i.e. syllables per utterances) it had a positive effect on *lexical* complexity (i.e. the total number of different words used).

4.2.1.3 Measure of Textual Lexical Density (MTLD)

Another *lexical* complexity measure employed in this study was the Measure of Textual Lexical Density (MTLD) Mean scores for each participant over the six tasks are displayed in Table 52 and graphically in Figure 23. The highest mean MTLD score was recorded for the LOW complexity FTF task. The lowest mean score was recorded for the HIGH complexity VW task. There is a general trend for the lower complexity tasks to have pushed learners to a higher lexical density. Additionally, although the mean MTLD scores for the two MID complexity tasks are comparable, in general, the FTF mean scores are higher than the VW equivalents.

Table 52: Mean scores and SD for MTLD.

	Mean	Std. Deviation
FTF LOW	15.91	6.82
FTF MID	10.61	3.23
FTF HIGH	10.90	4.52
VW LOW	12.16	5.50
VW MID	10.86	4.02
VW HIGH	8.96	3.77

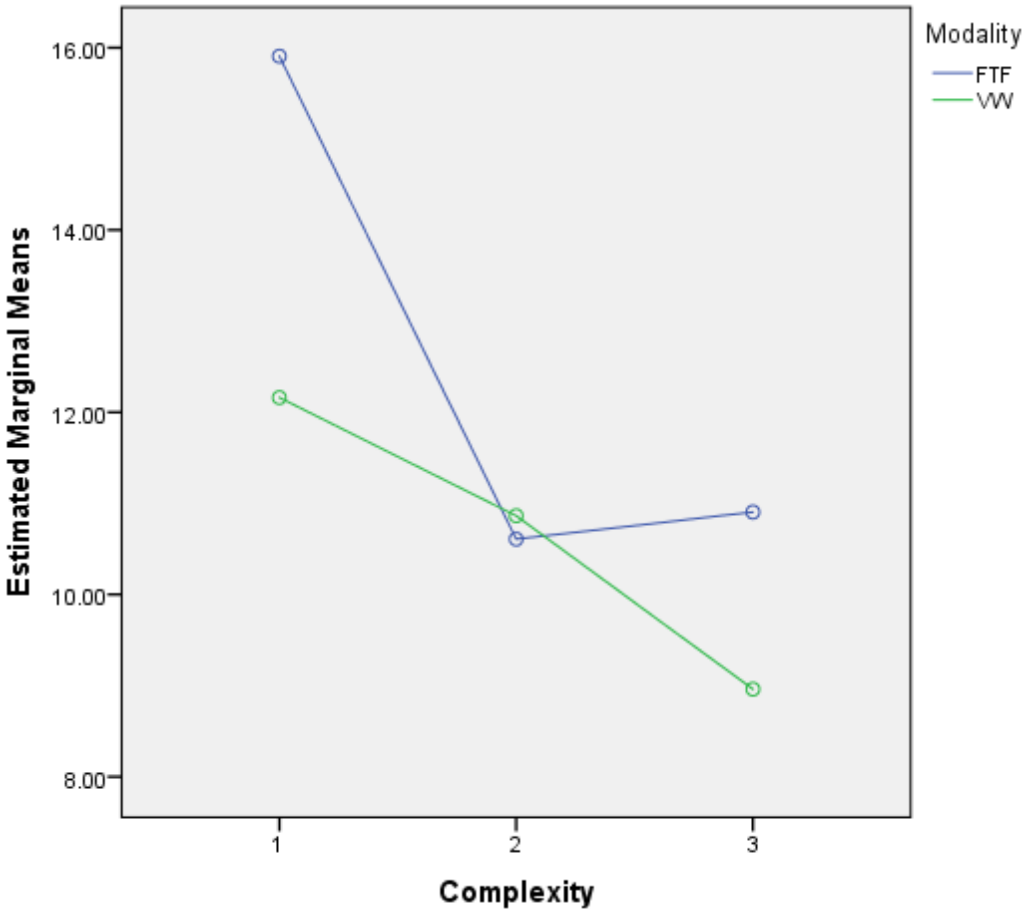


Figure 23: Estimated marginal means of the MTLD measure for all six tasks

The data was only normally distributed for one of the six tasks, as assessed by Shapiro-Wilk's test of normality on the studentized residuals (Table 53). Two outliers were identified by examining studentized residuals (FTF HIGH, value 3.33 and VW HIGH, value

3.08). The outliers were both associated with the same participant, who, as previously stated had a higher proficiency in English than his peers. The outlier was therefore not due to a miscalculation or data entry error. The two-way repeated measures ANOVA ran twice, once with outliers included and again with them removed. There were no differences in terms of statistical significance between the two sets of data. Therefore the data with outliers were used for further analysis.

Table 53: Normality tests for the different words measure

Task	Statistic	Df	Sig.
FTF_LOW	.77	20	.000
VW LOW	.84	20	.004
FTF MID	.91	20	.05
VW MID	.84	20	.003
FTF HIGH	.79	20	.001
VW HIGH	.86	20	.01

A two-way repeated measures ANOVA was run to determine whether there was an interaction between modality and task complexity for MTLD mean scores. Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for the two-way interaction, χ^2 (2) = 7.12, p = .03. Greenhouse-Geisser was used to correct the results. Inspection of the results revealed that there was a statistically significant two-way interaction, $F(1.51, 28.65) = 7.53, p < .05$. Additionally, in terms of main effects, there was also a statistical significance found for both modality and task complexity (see Table 54).

Table 54: Two-way repeated measures ANOVA for modality and task complexity on MTLD mean scores.

Source		Sum of squares	df	Mean Square	F	Sig.	Partial Eta Squared
Modality	Sphericity Assumed	98.39	1	98.39	18.68	.000	.5
Complexity	Sphericity Assumed	377.85	2	188.92	13.3	.000	.41
	Greenhouse-Geisser	377.85	1.64	230.39	13.3	.000	.41
Error (Complexity)	Sphericity Assumed	539.95	38	14.21			
	Greenhouse-Geisser	539.95	31.16	17.33			
Modality * Complexity	Sphericity Assumed	80.36	2	40.18	7.53	.002	.28
	Greenhouse-Geisser	80.36	1.51	53.30	7.53	.01	.28
Error (Modality * Complexity)	Sphericity Assumed	202.91	38	5.34			
	Greenhouse-Geisser	202.91	28.65	7.08			

Examining the main effect for modality via pairwise comparisons suggested that the difference in MTLD mean scores was 1.81 higher for the FTF mode over the VW mode, 95% CI [0.93 to 2.69], $p < 0.001$. For task complexity, the main effect revealed that participants had significantly higher MTLD mean scores when completing the LOW complexity tasks, compared to both MID and HIGH complexity tasks, regardless of modality. There was no statistical significance in mean scores for the MID and HIGH complexity tasks (Table 55). However, as the two way repeated measures ANOVA revealed an interaction between the factors, simple main effects were explored in more detail.

Table 55: Pairwise comparisons for task complexity main effect on MTLD mean scores.

Task Complexity	Task Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower Bound	Upper Bound
LOW	MID	3.30*	.99	.01	.71	5.88
	HIGH	4.10*	.87	.00	1.81	6.39
MID	HIGH	.80	.63	.66	-.86	2.47

Notes. *The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

4.2.1.3.1 Simple main effects of modality on MTLD scores

To determine the difference in mean MTLD scores in terms of modality, three comparisons of the data were conducted at the LOW, MID and HIGH levels of task complexity, results of which are presented in Table 56. Results revealed a statistically significant simple main effect for modality when completing the LOW and HIGH complexity tasks, but as can be expected from the analysis of the descriptive statistics, not at the MID complexity level.

Table 56: Pairwise comparisons for simple main effects of modality on MTLD mean scores

Task Complexity	(I) Mode	(J) Mode	Mean Difference (I-J)	Std. Error	Sig[a]	95% CI[a]	
						Lower Bound	Upper Bound
Low	FTF	VW	3.75*	.65	.000	2.38	5.11
Mid	FTF	VW	-0.26	.89	.78	-2.11	1.6
High	FTF	VW	1.94*	.62	.01	0.65	3.24

Notes. *The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

For this complexity measure, results suggest that modality had a significant effect on participants' performance. They performed better when doing the FTF tasks where mean MTLD scores for the LOW and HIGH complexity tasks were significantly higher in the FTF mode over the VW mode. However, the mean scores recorded for the MID complexity tasks did not show any statistically significant difference, where one would also expect the FTF task to be higher than the VW task based on the results for the LOW and HIGH complexity

levels. Reasons for this discrepancy will be explored by examining the qualitative data (transcribed data and responses to the open-ended questions of the questionnaires). One assumption is that specific task conditions of the VW MID task contributed to participants' poor performance along this dimension of output complexity.

4.2.1.3.2 Simple main effects of task complexity on MTLD scores

For the FTF tasks, the assumption of sphericity was violated, χ^2 (2) = 6.14, $p < .05$. Results suggest that there was a statistically significant effect of task complexity on participants' task performance in terms of syllables per utterances when completing the FTF tasks, $F(1.55, 29.4) = 14.62$, p = < .001 (Greenhouse-Geisser corrected). Subsequently, a significant difference in mean scores was found at the LOW-MID and LOW-HIGH levels but not at the MID-HIGH level (Table 57).

Table 57: Pairwise comparisons for the simple main effect of task complexity on MTLD scores for FTF modality.

(I) Complexity	(J) Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower	Upper
LOW	MID	5.3	1.35	.003	1.75	8.85
	HIGH	5.00	1.05	.000	2.26	7.75
MID	HIGH	-.3	.85	1.00	-2.52	1.93

Notes. *The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

For the one-way repeated measures ANOVA on the MTLD scores for the VW tasks, sphericity was met, χ^2 (2) = .86, p = .65 and a significant effect of task complexity on syllables per utterance was found, F(2, 38) = 6.98, p = .003. Examination of pairwise comparisons revealed that the only statistically significant difference in mean scores was found between the LOW and HIGH complexity tasks where there was a difference by 3.2, 95% CI [1.03, 5.37], $p < .0.5$.

Table 58: Pairwise comparisons for the simple main effect of task complexity on MTLD scores for VW modality.

Complexity	Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower	Upper
LOW	MID	1.30	.95	.56	-1.19	3.79
	HIGH	3.20*	.83	.003	1.03	5.37
MID	HIGH	1.90	.80	.09	-.20	4.01

Notes. *The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

In summary, then, results suggest that as complexity increased, there was a general reduction in learner output complexity, a result that seems opposed to the claims of the CH. The following section summarizes findings for all complexity measures.

4.2.1.4 *Summary for output complexity measures*

Table 59 provides a summary of simple main effects for modality on learners' output complexity.

Table 59: Summary of simple main effects of modality on output complexity

Measure		Task Complexity		
		LOW	MID	HIGH
Syll/Utt				
	Mean (FTF)	4.95	3.98	3.28
	Mean (VW)	4.03	3.58	3.09
	Mean diff (FTF – VW)	.92	.41	.21
	Sig.	.000	.02	.30
Different words				
	Mean (FTF)	59.48	67.45	58.13
	Mean (VW)	65.25	71	99.8
	Mean diff (FTF – VW)	-5.78	-3.55	-41.68
	Sig.	.11	.40	.00
MTLD				
	Mean (FTF)	15.91	10.61	10.90
	Mean (VW)	12.16	10.86	8.96
	Mean diff (FTF – VW)	3.75	-.25	1.94
	Sig.	.00	.78	.01

Results for the effect of modality and task complexity on learners' output complexity are as follows. For *structural* complexity, there was an interaction between modality and task complexity which had a negative impact on output complexity. Results for *lexical* complexity measures, however, were somewhat contradictory. Modality and task complexity interacted to produce a positive effect on learners' output complexity in terms of the mean number of different words they produced with the VW HIGH task pushing learners to the most complex

performance, however, the MTLD mean scores were somewhat contradictory to this, where mean scores for the VW tasks were lower than the FTF equivalents at the LOW and HIGH complexity level. Subsequently, a short summary of each measure is presented.

Learners produced significantly more *syllables per utterance* when performing the FTF tasks at both LOW and MID complexity levels, but not at the HIGH complexity level. One interpretation of this is that as cognitive demands imposed by communicating in the VW modality are reduced, learners were able to attend more of their attentional resources to their linguistic performance. This lends support to a trade-off effect and therefore the LACM conceptualization of how task complexity may affect task performance. The result is opposed to the claims of the CH.

For the *different words* measure, there were no significant differences in mean scores for the LOW and MID complexity tasks, which suggests that modality did not affect the number of different words that learners produced for these two task pairs. However, modality had a significantly large effect at the HIGH complexity level, where learners produced an average of 41.68 more different words when completing the VW task compared to the FTF equivalent. This suggests that the completing the task in the VW required learners to utilize more of their available vocabulary, and may be attributed to the task design which promoted learners to engage with specific game items, something that could be bypassed in the FTF mode. A more in-depth investigation of the discrepancy is carried out in the discussion chapter below.

MTLD mean scores were, on average, higher when learners completed the FTF tasks. This is true for the LOW and HIGH complexity tasks, where statistically significant differences in mean scores were found. The same is not found for the MID complexity task pair, where the two mean scores are almost the same. Results again suggest that the FTF mode may have allowed learners to focus more attention on their linguistic output, as attentional resources were made more available in the less cognitively demanding FTF mode. As a concrete example of the difference in performances recorded for this measure, the VW HIGH task promoted the worst performance from learners in terms of the MTLD measure where the mean score was only 8.96, a score significantly lower than the FTF equivalent at 10.9.

Interpreting this result then, it appears that the cognitive or spatial demands of the VW tasks may have required learners to repeat themselves in order for their interlocutor to understand their command, thus reducing the overall MTLD mean scores for this modality. This is because a low MTLD score indicates that the proportion of content words to the total number of tokens is low. As a concrete example, consider this excerpt of consecutive utterances from Participant 18:

> *"And Boss has a potion. And could you find... And looking at a pond. And still standing. And you? And you? And you? And you? Beeroman? Beeroman? Boss have a potion and he is sitting."*

The above, short excerpt has an overall lexical density score of 34.29%. Removing the repetitions of "And you?" and "Beeroman?" increases the overall lexical density score to 39.29%. The transcribed data is referenced in the discussion chapter to explore this hypothesis further.

4.2.2 Accuracy measure

This section provides results related to the accuracy measure employed in this study: the number of correct utterances. Following, each of the three error types (lexical, morphological and syntactic) are also analysed to assess whether modality had an effect on the type of errors that learners produced.

4.2.2.1 Correct Utterances

Mean scores for each task are displayed in Table 60 and visually in Figure 24. The highest mean scores were recorded for the MID complexity tasks, and the lowest mean scores were recorded for the HIGH complexity tasks for each mode of communication. The results of descriptive statistics for this measure reveal that there does not appear to be a main effect for modality. However, the next stage of analysis was to run a repeated measures ANOVA on the data.

Table 60: Mean scores and SD for the correct utterances measure.

Task	Mean (%)	SD
LOW FTF	51.7	25.80
MID FTF	68.8	10.59
HIGH FTF	36.8	18.09
LOW VW	52.1	19.54
MID VW	66.0	10.33
HIGH VW	46.8	20.37

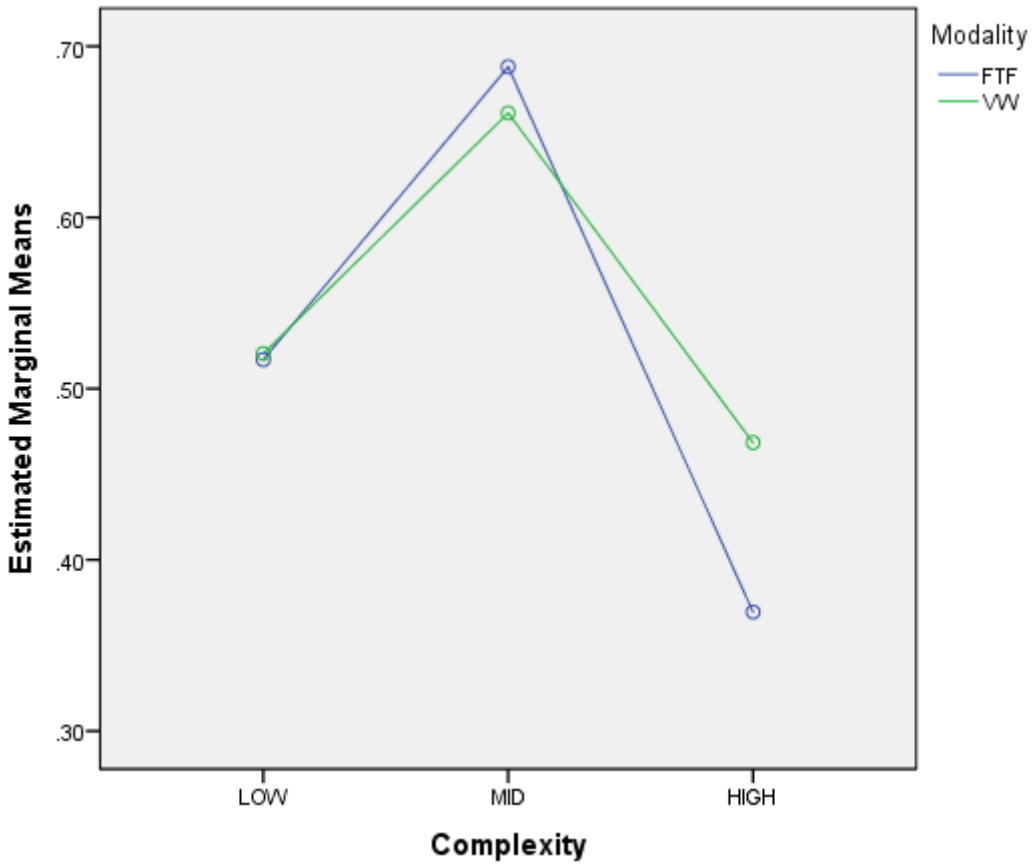


Figure 24: Estimated marginal means for the correct utterances measure for all six tasks

A two-way repeated measures ANOVA was run to determine whether there was an interaction between modality and task complexity on the mean number of error-free

utterances produced by learners. There was one outlier with a studentized residual of 3.05. This was recorded Participant 8's particularly accurate performance of the HIGH complexity FTF task (error-free utterances = 91%). As was previously found, this student's higher proficiency marked him as an outlier. The data was not removed. Subsequently, the number of error-free utterances produced by learners was normally distributed ($p < .05$) for all tasks as assessed by Shapiro-Wilk's test of normality on the studentized residuals.

Mauchly's test of sphericity indicated that the assumption of sphericity had not been violated for the two-way interaction, χ^2 (2) = .54, p = .77. Visual inspection of profile plots in Figure 24 indicated that a two-way interaction between modality and task complexity was unlikely as the lines for each mode were mostly parallel. Indeed, an inspection of the results of the two-way repeated measures ANOVA revealed no statistically significant two-way interaction, $F(2, 38) = 2.72$, $p = .08$. Additionally, there was no statistically significant main effect for modality, only task complexity $F(2, 38) = 25.86$, $p < .001$ (Table 61).

Table 61: Two-way repeated measures ANOVA for modality and task complexity on the volume of correct utterances.

Source	Sum of squares	df	Mean square	F	p	Partial Eta sq.
Mode	.02	1	.02	.93	.35	.05
Task Complexity	1.33	2	.66	25.86	.000	.58
Mode x Task Complexity	.09	2	.04	2.72	.08	.13

Examining the main effect of task complexity, significant differences in mean scores were found at the LOW-MID and MID-HIGH levels, where participant performances were significantly more accurate when completing the MID complexity task. There is a surprising result in that the mean accuracy scores at the LOW-HIGH level are not significantly different (

Table 62).

Table 62: Pairwise comparisons for task complexity main effect on the number of correct utterances.

Task Complexity	Task Complexity	Mean Difference	Std. Error	Sig[a]	95% CI[a]	
					Lower Bound	Upper Bound
LOW	MID	-.16	.04	.003	-.26	-.05
	HIGH	.10	.04	.06	-.002	.20
MID	LOW	.16	.04	.003	.05	.26
	HIGH	.26	.03	.000	.19	.33

Notes. *The mean difference is significant at the .05 level.

[a]Adjustment for multiple comparisons: Bonferroni.

This suggests that task complexity manipulations did not influence output accuracy in any meaningful way. Based on the claims of the CH, as task complexity increases, so should the accuracy of learner output. However, for these three task pairs, regardless of modality, learners' output accuracy was not significantly higher when completing either the LOW or HIGH complexity tasks. Alternatively, *task conditions* (participation and participant factors) or *task difficulty* variables (affective and ability factors) may have been more influential on accuracy, as participants produced significantly more accurate utterances when completing the MID complexity task. Reasons for this are explored in the discussion chapter below with reference to the speech and questionnaire data qualitatively.

4.2.2.2 Erroneous utterances

This section focuses on errors with the aim of exploring whether modality affected the type of errors participants made. Mean scores for all error types are displayed in Table 63. Graphical representations of the results are also available in Figure 25 (lexical errors), Figure 26 (morphological errors) and Figure 27 (syntactic errors). Inspection of mean scores, it is difficult to make any conclusions regarding how either modality or task complexity influenced error production.

Table 63: Mean scores and standard deviations for all error types.

Task	Lexical errors		Morphological errors		Syntactic errors	
	Mean (%)	SD	Mean (%)	SD	Mean (%)	SD
FTF LOW	43.60	24.49	16.05	17.43	5.95	11.80
FTF MID	29.00	10.28	3.40	3.95	4.10	5.73
FTF HIGH	56.50	19.01	9.70	7.85	8.20	7.21
VW LOW	41.70	19.11	11.90	8.96	7.40	10.55
VW MID	29.45	10.35	4.80	4.07	3.30	3.18
VW HIGH	48.75	20.37	6.70	6.51	9.15	8.28

For lexical and morphological errors, mean scores suggest that participants seemed to perform better in the VW mode, where mean scores are generally lower for these tasks in a task pair. However, the opposite is seen for syntactic errors, where mean scores are generally lower for the FTF tasks (a low score for errors indicates a more accurate performance). Additionally, one area of commonality for each of the measures is that there is a discrepancy between the LOW and HIGH complexity tasks and the MID complexity task. For instance, mean scores for morphological errors at the MID complexity level are lower for the FTF mode, whereas, at the LOW and HIGH complexity levels, mean scores are higher for the FTF mode. The same phenomenon is seen with syntactic errors but with FTF and VW task performances switched.

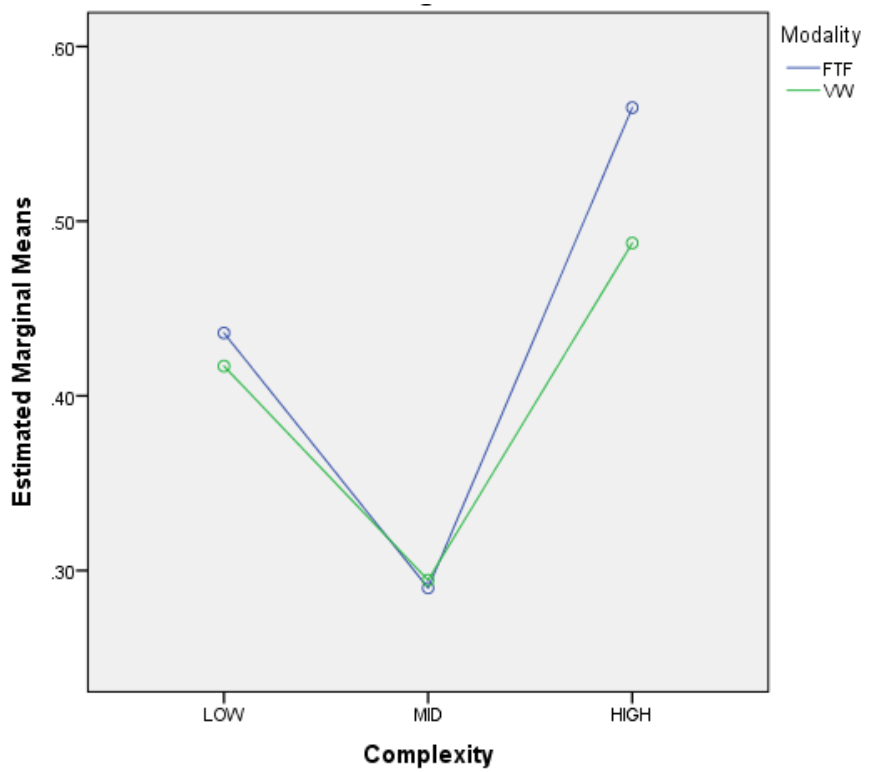


Figure 25: Estimated marginal means for lexical errors

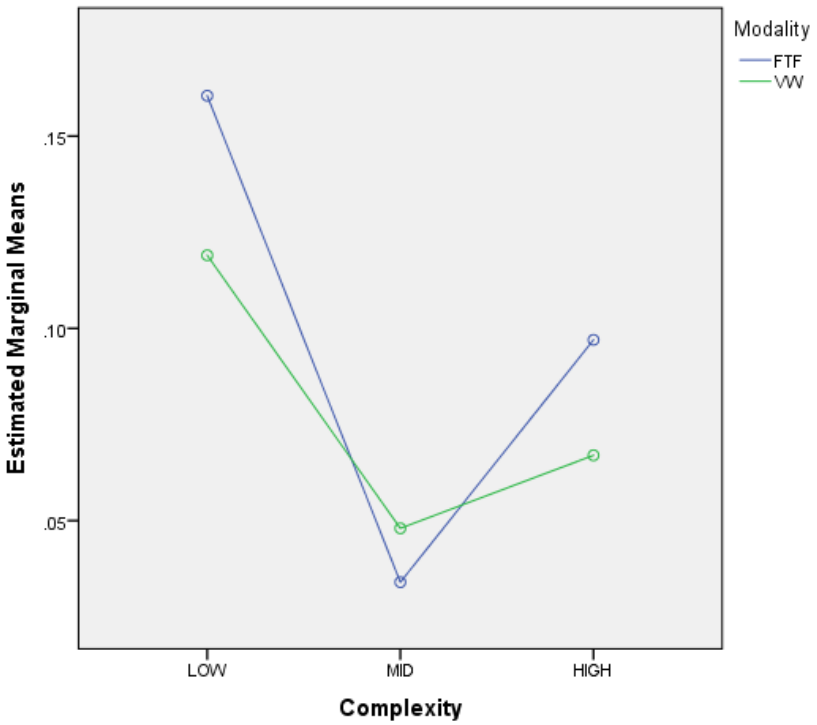


Figure 26: Estimated marginal means for morphological errors

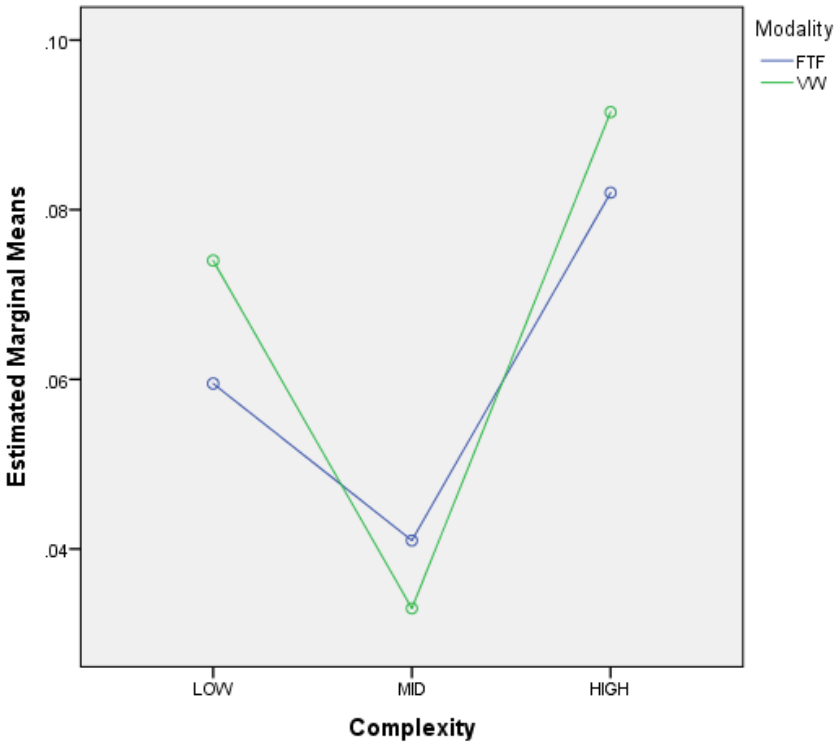


Figure 27: Estimated marginal means for syntactic errors

Following the descriptive statistical analysis, a two-way repeated measures ANOVA was ran on the data. For lexical errors, there were no outliers recorded for any of the six tasks as assessed by examination of studentized residuals for values greater than ±3 and the data were normally distributed, as assessed by Shapiro-Wilk's test of normality on the studentized residuals ($p > .05$). However, for morphological and syntactic errors, the data for a number of tasks was not normally distributed.

Mauchly's test of sphericity indicated that the assumption of sphericity was met for a two-way interaction between modality and task complexity for lexical and syntactic errors, but not morphological errors, where sphericity was violated. Subsequently, results of a two-way repeated measures ANOVA revealed that there was no two-way interaction between modality and task complexity for any of the error measures (Table 64). As a result, simple main effects were not investigated further. Instead, the main effects for modality and task complexity were referenced separately. Results for modality showed no statistically significant differences for any of the three error measures (Table 65), indicating that modality had no effect on the number of lexical, morphological or syntactic errors that participants made during task performance.

Table 64: Error measures: Within-subject effects for modality and task complexity

Measure	SS	Df	Mean Square	F	Sig.	Partial Eta Squared
Lexical Errors	0.04	2	0.02	1.14	0.33	0.06
Morphological Errors	0.02	1.51*	0.01	1.32	0.28	0.07
Syntactic Errors	0.003	2	0.001	0.32	0.73	0.02

*Greenhouse-Geisser corrected

Table 65: Main effects for modality on erroneous utterances

			95%CI	
Measure	Mean diff. (FTF – VW)	Sig.*	Lower	Upper
Lexical Errors	.03	.22	-.02	.08
Morphological Errors	.02	.18	-.009	.05
Syntactic Errors	-.01	.57	-.02	.01

*Adjustment for multiple comparisons: Bonferroni.

Regarding the main effect of task complexity on learners' errors, results echo those found for the correct utterances measure above. The only statistically significant differences in mean scores found were favourable towards the MID complexity task, where learners produced fewer errors when completing this task as opposed to the LOW and HIGH complexity tasks. A detailed overview of pairwise comparisons for the main effect of task complexity in learners' output errors can be seen in Table 66 below.

Table 66: Error measures: Task complexity main effects

Measure	(I) Complexity	(J) Complexity	Mean Difference (I-J)	Std. Error	Sig.[a]	95% CI[a]	
						Lower Bound	Upper Bound
Lexical Errors	LOW	MID	.13*	0.04	.01	0.03	0.24
		HIGH	-0.1	0.04	.07	-0.21	0.01
	MID	HIGH	-.23*	0.03	0	-0.32	-0.15
Morph Errors	LOW	MID	.10*	0.02	.001	0.04	0.16
		HIGH	0.06	0.02	.08	-0.01	0.12
	MID	HIGH	-.04*	0.01	.02	-0.08	-0.01
Syntactic Errors	LOW	MID	0.03	0.02	.7	-0.03	0.09
		HIGH	-0.02	0.02	1	-0.08	0.04
	MID	HIGH	-.05*	0.01	0	-0.08	-0.03

Based on estimated marginal means

* The mean difference is significant at the .05 level.

[a] Adjustment for multiple comparisons: Bonferroni.

In summary, descriptive analysis of error types revealed that lexical errors were the most common, with as much as 56% of total utterances containing a lexical error (HIGH complexity FTF task). Participants lexical errors also appeared normally distributed for all tasks; however, morphological and syntactic errors were not normally distributed. Two-way repeated measures ANOVAs revealed that modality had no statistically significant effect on learners' errors for all three error types, and task complexity main effects suggested that the MID complexity task produced the least amount of errors in learner speech. It is therefore hypothesized that task conditions other than inherent task complexity manipulations caused participants to make fewer errors during tasks of this task pair.

4.2.3 Fluency measure

Mean scores for each participant over the six tasks are displayed in Table 67 and graphically in Figure 19. Analysis of descriptive statistics indicates that for all task pairs, the

mean scores for tasks completed in the FTF mode were higher than the VW equivalents. The highest mean score was recorded for the LOW complexity tasks (FTF: 30.94 syll/minute, SD 10.71; VW: 19.60 syll/minute, SD = 11.40). The lowest mean scores were recorded for the HIGH complexity tasks for both modes of communication.

Table 67: Mean scores and standard deviations for syllables per minute

Task	Mean	SD
LOW FTF	30.94	10.71
MID FTF	30.39	9.79
HIGH FTF	23.75	10.84
LOW VW	19.60	11.40
MID VW	18.53	7.68
HIGH VW	15.42	9.93

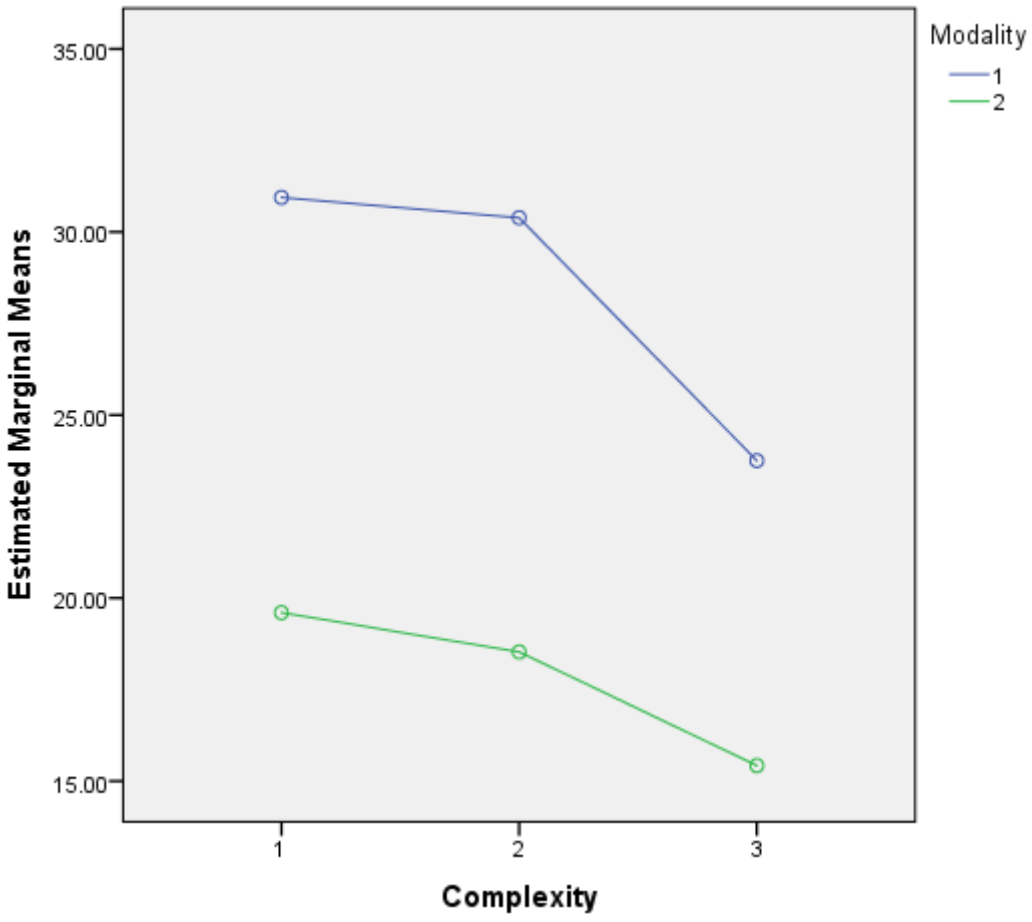


Figure 28: Estimated marginal means of the syllables per minute measure for all six tasks

A two-way repeated measures ANOVA was run to determine whether there was a two-way interaction between modality and task complexity on participants output fluency. There were no outliers, as assessed by examination of studentized residuals for values greater than ±3. Output fluency was normally distributed for four of the six tasks as assessed by Shapiro-Wilk's test of normality on the studentized residuals ($p > .05$). The two tasks that were not normally distributed were both the LOW complexity tasks. Participant 8 performed better than his peers for these two tasks. His scores were removed from the data, and the normality test ran again, which resulted in all data appearing normally distributed. The two-way repeated measures ANOVA was also ran with and without his scores, which resulted in no difference in statistical significance. The non-normally distributed data was therefore used to generate the following report.

Table 68: Shapiro-Wilk's test of normality on syllables per minute for each task.

Task	Statistic	Df	Sig.
FTF_LOW	.89	20	.02
FTF MID	.91	20	.05
FTF HIGH	.95	20	.29
VW LOW	.85	20	.01
VW MID	.98	20	.93
VW HIGH	.93	20	.15

Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for a two-way interaction between modality and task complexity, χ^2 (2) = 7.12, p = .03. Epsilon (ε) was 0.81, as calculated according to Greenhouse & Geisser (1959), and was used to correct the two-way repeated measures ANOVA. Visual inspection of profile plots in Figure 28 revealed that there was unlikely to be an interaction between modality and task complexity due to the lines appearing parallel and results of the two-way repeated measures ANOVA confirmed this. There was no statistically significant two-way interaction, F(1.51, 28.65) = .94, p = .38.

Examining the main effect for modality via pairwise comparisons suggested that the mean difference in the number of syllables per minute was 10.51 words higher in the FTF

mode over the VW mode, 95% CI [8.16 to 12.86], $p < .001$. For task complexity, the main effect revealed that the LOW complexity task allowed participants to produce statistically significantly more syllables per minute than the HIGH complexity tasks (5.68 syll/min., 95% CI [1.28 to 10.09], $p = 0.01$, but there was no significant difference between the LOW and MID complexity tasks (Table 69). Additionally, mean scores for the MID complexity tasks were also significantly higher than the HIGH complexity tasks.

Table 69: Pairwise comparisons for task complexity main effect on syllables per minute.

Task Complexity	**Task Complexity**	**Mean Difference**	**Std. Error**	**Sig[a]**	**95% CI[a]**	
					Lower Bound	Upper Bound
LOW	MID	.81	1.89	1.00	-4.15	5.78
	HIGH	5.68	1.68	.01	1.28	10.09
MID	HIGH	4.87	1.22	.002	1.67	8.07

Studies which investigate the effect of task complexity manipulations on output fluency have found either a negative relationship between fluency and task complexity (Michel, Kuiken, & Vedder, 2007; Robinson, Cadierno, & Shirai 2009; Levkina & Gilabert, 2012), or no effect on fluency (Gilabert, 2007; Révész, 2011; Sasayama & Izumi, 2012). Results of this study seem to support these findings, where fluency was hindered at the higher complexity level. Additionally, and more importantly, modality had a statistically significant effect on learner output fluency at all levels of task complexity, which suggests that modality may be more influential in determining learner output fluency than task complexity manipulations.

4.3 CHAPTER SUMMARY

This chapter presented and analysed the data from transcriptions and the post-task questionnaires in order to explore the differences in learner output as they completed three different task pairs. This was undertaken with the aim of answering the main research question. That is, how *modality* and *task complexity* affect learner oral task performance.

Statistical analyses of the transcription data revealed that *structural complexity* was negatively affected by modality and task complexity. The lower complexity FTF tasks

pushing learners to produce more syllables per utterance. There were mixed results for *lexical complexity* where the total number of different words participants spoke was higher when completing the VW and more complex tasks (an interaction of modality and task complexity). However, *MTLD* mean scores were lower when participants completed the VW tasks, suggesting that the increased cognitive demands of task completion in the VW hindered complex output. The fluency measure revealed a strong negative relationship between modality and fluent output where FTF tasks pushed learners to output 10.51 words per minute more than the VW tasks on average. A statistically significant finding. Accuracy was not affected by modality, and seemingly not by task complexity either. There was a main effect of task complexity on learners' output accuracy, but it was not as predicted. Learner output was most accurate for the MID complexity task which suggests that task complexity manipulations were not the primary influence on learner output.

Subsequently, the questionnaire data were analysed with the aim of answering RQ2, that is, how modality and task complexity affected learner attitudes towards studying English, and more specifically, what affective affordances of the VW. Based on the results of previous studies in the literature, it was hypothesized that the authentic and immersive properties of the VW would have a positive impact on learner motivation. Findings suggest that this was the case, and interestingly, although the VW tasks were perceived to be more difficult and mentally demanding than the FTF equivalents, they were considered to be more enjoyable.

5 DISCUSSION

5.1 INTRODUCTION

In this chapter, I compare the findings of this study with those of previous research in the literature, highlighting instances where results seem to coincide or differ, and provide further explanation as necessary. I also make reference to the transcribed speech data qualitatively in order to explore reasons for the results of the statistical tests. Finally, participants' responses to the open-ended questions on the two questionnaires are also referenced qualitatively in more depth. This data is explored in order to gain further insight into the possible reasons for any results.

5.2 THE RELATIONSHIP BETWEEN MODALITY, TASK CONDITIONS, AND ORAL TASK PERFORMANCE

The following sections explore the possible reasons for differences in participant oral task performance as they completed the online and offline tasks. Each of the three performance measures is considered sequentially, starting with complexity, then accuracy and finally fluency.

5.2.1 Output complexity

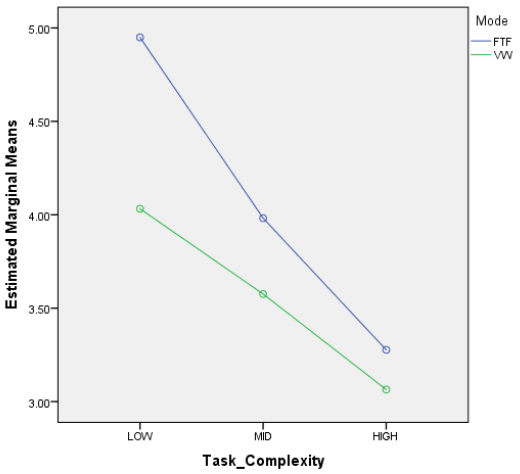
In the current study, an inverse relationship between task complexity and oral syntactic complexity was found. Comparing output complexity for the designed-to-be simplest task pair and the designed-to-be most complex task pair revealed that the interaction of task complexity and modality had a significantly negative effect on participants' performance for both the *syllables per minute* and *MTLD* measures. In terms of modality, it appears that FTF tasks used in this study either 1) pushed participants to produce more complex language compared to their VW equivalents, or, 2) that the cognitive demands of the simpler (FTF) tasks allowed them to focus their attention on producing more complex output. This finding is counterintuitive to the claims of the Cognition Hypothesis which posits that as task complexity increases, learner output should increase accordingly in order to match the demands of the task. Alternatively, then, this finding seems to lend support for the claims of the LACM. That is, it may be hypothesised that the additional cognitive demands of controlling avatars and traversing the virtual landscape had a negative effect on

the participants' ability to produce complex language due to a lack of attentional resources available for language production. Thus, a trade-off effect may have occurred.

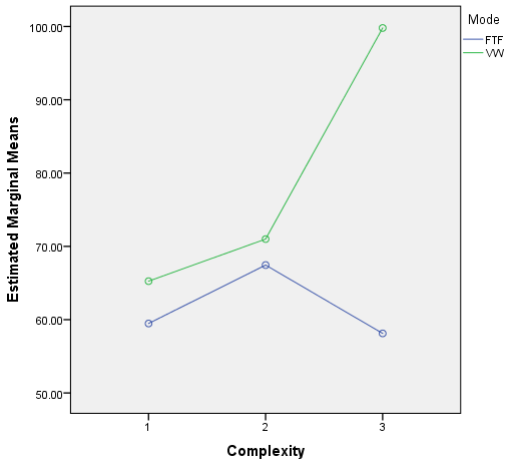
In a recent, related study by Sasayama (2015), it was found that of four monologic, narrative tasks, the most designed-to-be and measured-to-be (based on participant perceptions of task difficulty) cognitively demanding task did not promote participants to produce the most complex output. Instead, the measured-to-be second simplest task pushed them to produce more complex output. Sasayama suggests that the reduction in cognitive demands of the simpler task may have freed up attentional resources which learners could then devote to task performance instead. Although task design and task conditions differ significantly between Sasayama's study and those used in this dissertation, results seem to coincide. Both results lend support for the occurrence of a trade-off effect, and therefore the claims of the LACM, rather than the claims of the CH.

One implication of this finding is that task design for immersive online environments should take into account the possible additional cognitive demands that the environment places on learners. If controlling an avatar, manipulating and traversing the virtual terrain, and using SCMC software poses too great a demand on learners' attentional resources it may hinder the ability to produce the L2. Consequentially, and from an interactionist perspective to SLA, this implies that opportunities for successful language acquisition are also reduced. In conclusion, then, tasks designed to be carried out in VWs should be designed to be simpler than FTF equivalents to allow for the increased cognitive demands of the environment.

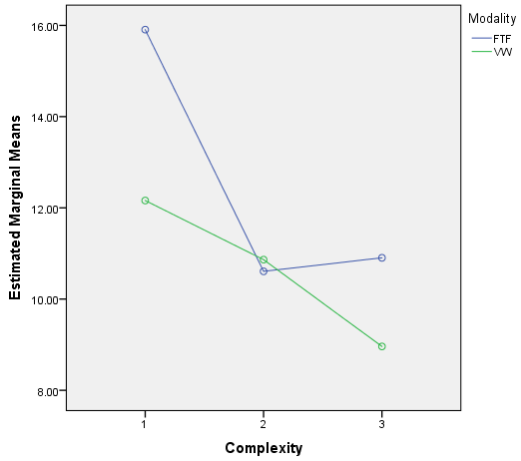
Secondly, results coincided for two of the measures used to assess output complexity. Syllables per utterance and the MTLD mean scores generally follow the same pattern with the LOW complexity tasks eliciting the most complex output. The *different words* measure, however, does not follow this pattern. See Figure 29 for a graphical representation of mean scores for all three complexity measures.



Syllables per utterance



Number of different words



MTLD

Figure 29: Graphical representation of mean scores for all complexity measures.

Mean scores for the MTLD measure suggest that the LOW complexity tasks promoted participants to produce the longest utterances whereas the HIGH complexity tasks promoted the shortest utterances. However, results for the *different words* measure are somewhat opposite to this, particularly when considering the main effect of modality. Additionally, it is only the *different words* measure that matches the original hypothesis of the CH: that task complexity should have a positive correlation with lexical complexity.

What, then, caused the differences in mean scores recorded for the MTLD and the different number of words measures? Although they are both concerned with lexical complexity, descriptive and statistical analyses revealed very different results for these two measures. I explore this topic in more detail in the following section.

5.2.1.1 Affordances of the VW for promoting lexically diverse output

Examining the breadth of vocabulary that learners were required to use during the six tasks may explain why the mean number of different words that participants used was so high for the VW HIGH task. This room decoration tasks were designed to require learners to use the same number of in-game (Minecraft specific) lexical items for both the VW and FTF version of the task. In comparison, with the MID complexity directions tasks, learners are generally relying on a fixed number of lexical items unrelated to the Minecraft in particular ("go straight down this road…"), for the room decoration tasks, however, participants were required to use game-related lexical items. The addition of these specific, Minecraft-related lexical items may account for the high mean score for the *different words* measure recorded for the Room VW task. However, one question regarding this remains unexplained. Results showed that for the VW HIGH task participants produced on average 41 (95% CI, 25.85 to 57.50) additional different words than the FTF task, $F(1, 19) = 30.37$, $p < .05$). What caused there to be a statistically significant difference in mean scores between the two HIGH complexity (room decoration) tasks?

The answer to this seems to come from an overlooked factor regarding the number of unique items that each task required participants to interact with. Focusing specifically on the HIGH complexity tasks: although care was taken to ensure that the number of elements in this task pair was as close as possible, it appears that participants were able to use circumlocution as a communication strategy when completing the FTF task but not the VW task, thus allowing them to bypass the use of game-specific vocabulary for the FTF version. That is, placing items in the VW required participants to make reference to the many specific items in the game and choose precisely the correct item in order to complete the task successfully. With the HIGH complexity FTF task, however, participants could draw a generic "table" where the diagram may have featured a table made of a *crafting table* block, an *oak*, *birch*, or *spruce log* block, or a *stone half slab* block among others (all game-specific items). This detail could easily be overlooked with the FTF task as participants did not need to select the exact block. However, when completing the VW equivalent, participants would have to explicitly state which block was needed to be placed, find it in their inventory and place it, thus promoting them to use the specific vocabulary and therefore increasing the total number of different words they used as part of task completion.

Figure 30: Screenshot of the Room FTF bedroom recreation task highlighting the items participants were required to recreate.

Introducing a concrete example of this is with the HIGH complexity FTF task. In this task, participants were required to recreate a bedroom scene as depicted in Figure 30. One participant would provide instructions to their interlocutor in order for them to draw it on their worksheet. In the picture, to the left and right of the bed were two items which look like pictures or paintings of a helmet and sword respectively. These objects could quickly be redrawn without directly having to reference the Minecraft-specific items that they are made of. That is, both "pictures" were comprised of a number of particular game-related items which would need to be chosen appropriately if the task was carried out in the VW. These items are *item frames* which contain an *iron helmet* (as opposed to *leather, gold* or *diamond helmet*) and an *iron sword* (as opposed to a *wooden*, *gold* or *diamond* sword). However, none of the participants mentioned these items when completing the FTF version of this task. Instead, the most common description was of a helmet or sword "picture" as can be seen in lines 272 and 273 in the following excerpt (Table 70):

Table 70: Excerpt of Participant 19 and 20 completing the Room FTF task

Utterance Number	Participant	Utterance
271	P19	Painting is here.
272	P19	Ah right side, there is sword picture
273	P20	Sword picture?
274	P20	Right side?
275	P19	Right side.
276	P20	Umm... Right side?
277	P20	Side?
278	P19	On the wall.
279	P19	On the wall.
280	P19	Up side.
281	P20	Beside painting.

Additionally, manipulating of blocks around their axes (both vertical and horizontal rotation is possible), as well as the various additional block attributes such as "on" or "off" for switches, "normal" or "upside down" for stair blocks, and other such unforeseen attributes may have affected how many different words participants spoke. For instance, Figure 31 shows a "*stone stairs block"* in both normal (left) and inverted (right) orientation. This apparent roadblock to successful communication can be considered an affordance of VWs in that learners can be pushed to interact with specific vocabulary more than the FTF mode. That is, learners are not able to bypass using specific vocabulary when undertaking the VW tasks as they can with FTF communication.

Figure 31: Stone stairs block orientation in Minecraft.

5.2.1.2 VW tasks afforded negotiation for meaning at the expense of output complexity

In direct opposition to the above finding, statistical analysis of participants' oral suggested that modality may have had a negative effect on output complexity in terms of lexical density and the number of syllables per utterance. This finding is somewhat counterintuitive to the Cognition Hypothesis and is explored in detail in this section.

In terms of the MTLD mean scores, participants' repetitive output may have been a contributing factor. Looking qualitatively at the transcription data, and in particular the data for the participant who produced the most number of utterances (Participant 08), it appears that the VW HIGH task was complex in a way that required a great deal of negotiation for meaning between the participants as they experienced communication breakdown. Participants had to repeat what they were saying, or reformulate their utterances in order to portray a correct instruction to their partner. An example is provided below in Table 71. In this excerpt from Participant 07 and 08, Participant 08 talks for 23 lines in order to get his partner to place a particular block in a particular position. This sample of text has an MTLD value of 23.50.

Table 71: An example of participants' repetitive output in the VW HIGH task

Utterance number	Participant Number	Utterance
256	08	I want you to put a lever above the cauldron.
257	08	I want you to put it to the wall.
258	08	You know.
259	08	So that it looks like a tap.
260	08	That's not right.
261	08	I said, please go to the right.
262	08	The right side.
263	08	How do I call it?
264	08	One block to the left from the very the limit of the right side of the room
265	08	I don't know how to call it.
266	08	The entrance.
267	08	The limit of the entrance.
268	08	I want you to go one block to the right
269	08	To the right for you
270	08	Sorry
271	08	Yeah, that's it.
272	08	Could you please make the lever?
273	08	I mean…
274	08	Bend the lever to…
275	08	I don't know how to...
276	08	OK, please put the lever in the same place.
278	08	And right click it.
279	08	Yeah, that's it.

Compare this with the same number of utterances spoken by this participant during the FTF LOW complexity task in Table 72. Participant 08, making himself understood the first time, produces output with a much higher MTLD value (49.02) for the exchange.

Table 72: An example of clear, precise output in the FTF LOW task

Utterance Number	Participant	Utterance
13	P08	He's enjoying?
14	P07	Yeah, enjoying.
15	P08	So, I guess my Beeroman is different from your Beeroman.
16	P07	Yeah, I think so.
17	P08	Roger that.
18	P08	So, what is Boss doing in your picture?
19	P07	In my picture, he is throwing a ball or playing billiards.
20	P07	I'm not sure of the pronunciation
21	P08	Billiards?
22	P08	I don't know the pronunciation either.
23	P08	In my picture, he is doing billiards
24	P08	So, our Boss is the same
25	P07	Yeah, I think so.
26	P08	Roger
27	P08	Sorry, give me a minute.
28	P07	OK.
29	P08	OK, go ahead.
30	P07	In your picture, what Cheapshot is doing?
31	P08	Cheapshot is cooking a fish or burning a fish; I don't know.
32	P08	Something like that.
33	P07	It's different from in my picture.
34	P07	In my picture, he is trying to kill zombies.
35	P08	Then I guess it's different.

One further, an extended example of this can be seen in Table 73 below. The table contains an excerpt of Participant 15 and 16 completing the VW HIGH complexity task. The VW task presented numerous technical problems for the participants. First, the terminology used for specific items was unclear to Participant 15 who uses the word “switch” instead of

the actual in-game term “lever” on line 119 (see Figure 32 for reference).

Table 73: An example of Participant 15 and 16 completing the Room VW task

Utterance Number	Participant Number	Utterance
117	P15	OK.
118	P15	So, next….
119	P15	You have switch.
120	P16	Switch.
121	P16	Switch block?
122	P15	Switch.
123	P16	Lever?
124	P15	Lever.
125	P15	Lever.
126	P15	OK.
127	P15	Redstone...
128	P15	Wait.
129	P16	OK.
130	P15	Under the Redstone lamp.
131	P16	Under the Redstone lamp.
132	P16	No distance?
133	P15	No.
134	P15	Put under the Redstone lamp.
135	P16	Not sure?
136	P15	Yeah.
137	P15	No.
138	P15	Under the Redstone lamp.
139	P15	OK.
140	P16	On the Redstone lamp?
141	P15	OK, so wait.
142	P15	OK.

Following, in lines 130 to 138, Participant 15 did not know how to successfully

instruct his partner to attach the lever to the bottom of the *redstone lamp* item, causing considerable confusion. Participant 16 finally understood what his partner was requesting and confirmed this on line 140. This type of communication breakdown was typical for this task as participants struggled to verbalise instructions for their partners. The lack of salience in terms of how to correctly navigate and manipulate the environment could have had a negative effect on output complexity, as well as fluency, as participants required extended periods of time to formulate an instruction, and had to wait for their partner to both interpret and act upon it. Based on the comparison made between the VW and FTF task exchanges above, I propose that the difference in MTLD mean scores for the two modes is based on the following affordances of the VW.

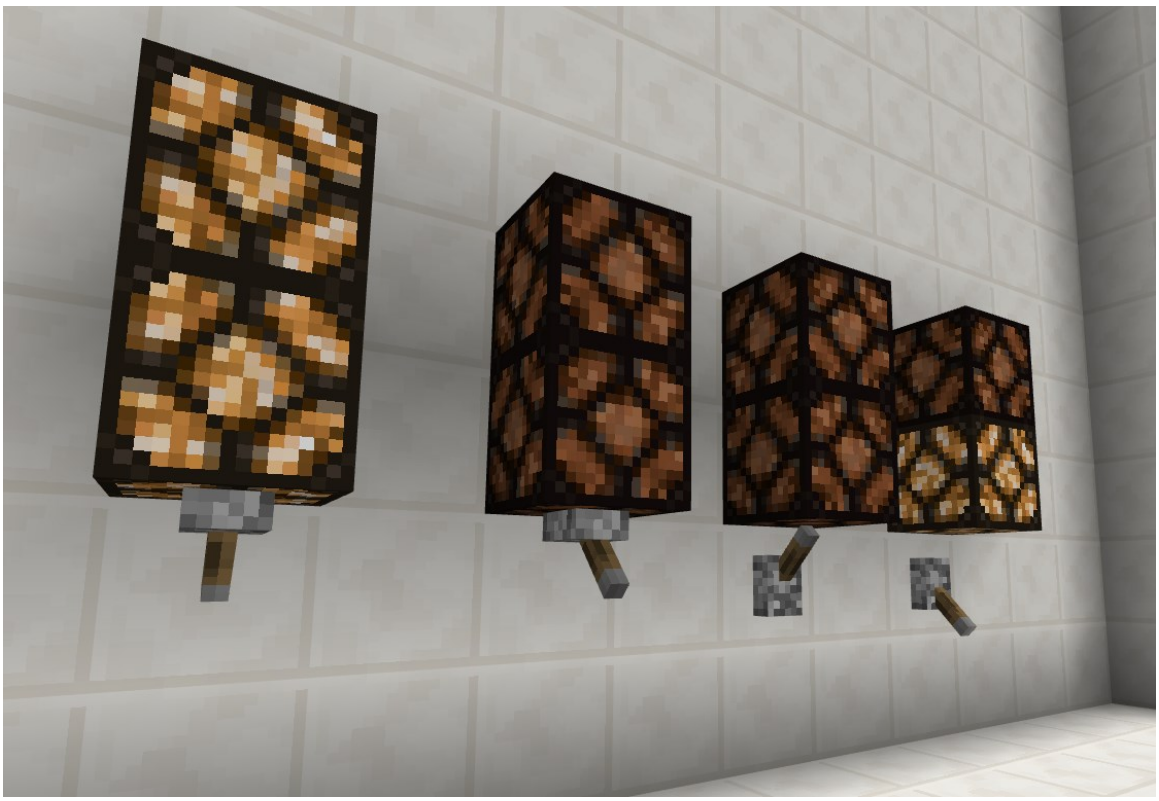


Figure 32: Redstone lamps and torches in the VW.

From left to right:
Redstone lamp with the lever in the ON position attached to the bottom. Only this configuration is acceptable.
Redstone lamp with the lever in the OFF position attached to the bottom.
Redstone lamp with the lever in the OFF position attached to the wall.
Redstone lamp with the lever in the ON position attached to the wall.

Although not investigated in any formal depth with statistical analyses, one positive result highlighted by the above excerpt is that the VW tasks appeared to promote participants

to engage in more *negotiation for meaning* than the FTF tasks, which from an interactionist model of SLA is a precursor to interlanguage development as learners switch their attention from meaning to form. This negotiation for meaning occurred due to the particular challenges afforded by the environment such as navigating the terrain, controlling an avatar, technical issues involving the communication software, and the complex mechanics of the environment such as learning the rules which govern how blocks are positioned. However, lexical diversity was reduced because of these challenges because engaging in negotiation for meaning meant that participants repeated the same instruction multiple times, reducing overall lexical density.

5.2.1.3 Lack of visual information and its effect on communication breakdowns

With reference to transcriptions of participants' task performance, the above sections revealed that the VW tasks promoted communication breakdowns and language repetition. One more reason for such breakdowns may be attributed to the lack of visual information (or at least, the lack of salience regarding an interlocutor's orientation to the environment) presented to participants in the VW. Table 74 below contains an excerpt from Participants 19 and 20 as they undertake the VW MID complexity (directions) task. Due to the cognitive demand of having to pay attention to the VW from their own, subjective viewpoint and guide their partner at the same time, it appears that participants experienced difficulties in establishing which direction their partner was facing. Therefore, providing accurate information regarding the direction they should move was problematic. After multiple attempts to get his partner to walk in a particular direction, Participant 20's commands seem to have failed, which is signalled by lines 172 and 173. As can be seen in the excerpt, participant 20 still did not manage to express his intentions correctly to Participant 19 even as late as Line 180.

Table 74: An example of repetition due to difficulties of using the VW

Utterance Number	Participant Number	Utterance
166	P20	Walk all the way down the road.
167	P20	Walk all the way down the road.
168	P20	Walk all the way down.
169	P19	Down?
170	P20	Walk all the way down.
171	P20	Walk all the way down this road.
172	P20	Hey, where are you?
173	P20	Where are you going?
174	P20	Walk down.
175	P19	Turn right?
176	P20	What!?
177	P19	Turn straight?
178	P19	Straight?
179	P20	Walk down!
180	P20	Walk all the way down this road.

Similar problems do not occur when the pair complete the FTF version of the same task. Table 75 below highlights how easily commands were understood in the FTF version of this task. As the pair both have precisely the same bird's eye view of the map, Participant 20 is able to give more extended commands knowing that his partner will understand where to go. This may have led to greater syntactic complexity in terms of the words per utterance measure, at least for Participant 20 in this case. Additionally, in the excerpt below, the output generated by Participant 19 is limited to giving confirmations of having followed the directions correctly with "OK." However, in exchanging roles later in the task, it is safe to assume that average complexity for the pair was high as Participant 19 had the opportunity to give directions and produce similarly complex utterances.

Table 75: An example of the same participant pair completing the FTF version of the same

directions task.

Utterance Number	Participant Number	Utterance
86	P20	First, walk down this street
87	P20	And turn left just after the first house.
88	P19	OK
89	P19	Next
90	P20	Go straight through the trees and out onto the road.
91	P19	OK
92	P20	Keep going straight up the steps.
93	P19	OK
94	P20	Can you see the boat over to your left?
95	P19	OK
96	P20	Keep walking straight and go up some more steps.
97	P19	OK
98	P20	Next, In front of you is a bridge
99	P20	Please cross the bridge.
100	P20	Cross.
101	P20	After you have crossed the bridge, immediately turn left.
102	P19	Turn left?
103	P20	Left.

5.2.2 Output accuracy

There was no two-way interaction between modality and task complexity for this measure as revealed by a two-way repeated measures ANOVA. There was also no statistically significant main effect of modality on the number of correct utterances produced by participants. There was however a statistically significant main effect for task complexity on accuracy where it was revealed that the participants were able to produce statistically significantly more correct utterances when completing the MID complexity tasks than both the LOW and HIGH complexity tasks. Task complexity manipulations, therefore, did not seem to affect participant output in terms of accuracy.

Whilst there are no studies that directly compare oral task performances between VW-based, and FTF modes, Baralt (2013) examined the effects of simple and complex tasks on participants' oral interactions in FTF and SCMC contexts. Their conclusion was that greater task complexity led to increased L2 development in the FTF context, but simple tasks led to better results in the SCMC context. In other words, there was an incongruity between the two modes of communication focused upon; task complexity improving performance in FTF contexts but hindering performance in SCMC contexts. Similarly, Nik (2010) found that simple tasks used in a written SCMC context led to improved accuracy but not complexity, suggesting that increased task complexity may not facilitate L2 learning in SCMC contexts in the same way that it does in FTF contexts, at least in terms of accuracy. Findings here do not coincide with those of Nik or Baralt in that accuracy was not affected significantly by modality.

Task complexity appeared to be a contributing factor in determining participant output accuracy, but again, it was not as expected. The highest mean scores being recorded for the MID complexity tasks. I argue that the resource-dispersing task complexity (+/-) *task structure* may have been particularly influential in determining this result. This is because the linguistic requirements of the MID complexity (directions) tasks are the most formulaic of all three task pairs. In other words, the language used in the MID complexity tasks may be reduced to a set of only a few, fixed phrases, or "language chunks" (see Ellis, 2005) such as:

- At the [ordinal number] corner, turn [direction].
- Go [up/down] this street.
- Go straight and turn [direction] at the [ordinal number] corner.

Unless there is a breakdown in communication (i.e. a learner mishears a direction and takes an incorrect path), participants were not required to step outside of this tight linguistic boundary and may complete the activity by using only a few formulaic language chunks. This point relates to one of Skehan's generalizations regarding the claims of the LACM in that tasks based on familiar information may promote increased accuracy and fluency (2001).

Looking at participant utterances qualitatively, this appears to be the case as

exemplified in Table 76. This excerpt also reveals a lack of communication breakdowns and therefore the absence of any negotiation for meaning episodes.

Table 76: Excerpt of Participant 03 and 04 completing the FTF Directions task.

Utterance number	Participant Number	Utterance
24	P04	Where is big red's house?
25	P03	Big Red's house.
26	P03	OK.
27	P03	Walk to the…
28	P03	Go down this street.
29	P03	And turn left at the corner before the church.
30	P04	OK.
31	P03	First corner turn left.
32	P03	Go down the steps.
33	P04	OK.
34	P03	Second corner turn right.
35	P03	Yes.
36	P04	OK.
37	P03	Go straight.

Additionally, as mentioned above in the discussion of the effect of modality and task complexity on participants' output complexity, the directions tasks did not require them to make explicit reference to game-related vocabulary. Directions could be given using more generic vocabulary such as "church." This may have had the effect of increasing participants' (+/-) familiarity with the task content and reducing the overall cognitive demands of the task.

Referencing participant responses to the open questions, it appears that the FTF MID task provided little new learning material, especially regarding new vocabulary. This backs up the claim that learner familiarity with task contents may be an influential factor in determining performance. Examples are provided below:

Participant 2 (Directions FTF comment): Alley と言う単語を学んだ。最初は valley のスペルミスだと思った。[I learnt the word 'alley.' I thought it was a misspelling

of ‘valley’ at first.]

Participant 7 (Directions FTF comment): Alley と pier と言う単語を学んだ。 [*I learnt the words "alley" and "pier."*]

Participant 20 (Directions VW comment): この活動で新しい事は学ばなかった。 [*I didn't learn anything new in this activity.*]

In summary, participant familiarity with the language used in the MID complexity tasks may have had a positive effect on accuracy during task performance rather than manipulations of resource-directing task complexity conditions.

5.2.3 Output fluency

There was no statistically significant two-way interaction between modality and task complexity for this measure. There was however a statistically significant main effect of both modality and task complexity. In terms of modality, output was statistically significantly less fluent when completing the VW tasks. One reason for this may be attributed to the (+/-) *here-and-now* task complexity variable. Although participants were co-located in the same room, their interactions were mediated via the technology (computers) between them, and their actions were not carried out by physically, but through the manipulation of an avatar in a (remote) 3D environment. The VW tasks thus represented a decrease in the here-and-now task condition. As hypothesised by the CH and found in studies that manipulate this task condition (Rahimpour, 1999; Robinson, 1995), a manipulation of the *here-and-now* task condition may have a negative effect on fluency.

However, and more obviously, actions in the VW required more time to complete than the FTF tasks. One reason for a decrease in output fluency when participants completed VW tasks may be attributed to the amount of time required to complete actions in this environment. Actions in the VW included traversing (walking through) the virtual landscape, manipulating objects, and controlling avatars. A more thorough list of example actions is provided in Table 77. The table lists common actions performed in this study, comparing VW and FTF modes where applicable.

Table 77: Comparison of common actions in the VW and FTF equivalents

Action in the virtual world	Real world (FTF task) equivalents
[DIRECTIONS VW] Walk an avatar down a virtual street.	Trace a street on the paper with a finger. Visually trace the street.
[ROOM VW] Choose the correct item from the inventory of items.	No equivalent. Players can draw the item directly.
[ROOM VW] Place blocks in the correct orientation. If the block is in the incorrect orientation, it must be destroyed and replaced.	No equivalent. Players are not restricted by the virtual world rules regarding block placement.
[SPOT VW] Navigate an avatar through a virtual house (into rooms, go upstairs, downstairs, outside, etc.) to search for a character.	Look over the worksheet diagram until the character is found.

It is clear that the VW tasks required participants to engage in non-verbal actions to a greater degree than the FTF tasks. Additionally, these actions also required more time to complete, thus longer stretches of time where participants were not speaking, but waiting for their interlocutor to complete an action. Therefore, the current study reveals that for tasks which require learners to manipulate an avatar, traverse terrain and complete actions in virtual environments, such activity hinders output fluency. As a concrete example, consider the two excerpts below from the same two participants as they complete the FTF LOW and VW LOW tasks. Both excerpts are 60 seconds in length (Table 78 and Table 79).

Table 78: First 60 seconds of audio for Participant 9 and 10 completing the FTF LOW task

Utterance number	Participant Number	Utterance
1	P09	In your picture, what is Beeroman doing?
2	P10	He is listen to music
3	P10	How about you?
4	P09	He is driving a car.
5	P10	Different.
6	P09	All right.
7	P09	In your picture, what is the boss doing?
8	P10	He is playing billiards.
9	P09	Me, too.
10	P10	OK, same.
11	P09	In my picture, Cheapshot is eating bread and fish.

Table 79: First 60 seconds of audio for Participant 9 and 10 completing the VW LOW task

Utterance number	Participant Number	Utterance
1	P10	OK, and him…
2	P10	He is drinking potion
3	P09	Potion?
4	P10	Drinking potion in dining.
5	P09	OK, OK.
6	P10	Beeroman.
7	P10	Beeroman.
8	P10	Beeroman.

When the participants complete the FTF LOW task there is minimal pausing between utterances as they move from one character to the next. But for the VW equivalent, there are considerably longer pauses as they have to move their avatar through the environment to search for characters. Indeed, the conversation presented in the VW excerpt suggests that Participant 10 has not explained which character he is talking about, causing him to repeat

the character's name "Beeroman" three times as he waits for his partner to locate the same character. This misunderstanding between the two participants is perhaps due to a lack of visual clues in the VW.

At the end of the 60 second period, Participant 09 had still not been able to locate this character, and so the two participants are unable to confirm whether their individual instances of "Beeroman" are undertaking the same activity or not. Incidentally, in this example, the two characters are not undertaking the same activity. Beeroman is on the first floor of one participant's house but on the second floor of the other participant's (see Figure 33). Having to find a character before being able to talk about its activity can be seen to have a negative effect on output fluency in this example.

Figure 33: Beeroman's location in the two Spot VW houses.

5.3 TASK CONDITIONS THAT INFLUENCED TASK DIFFICULTY PERCEPTIONS

Following, this section will explore which task conditions contributed to participants' perceptions of task difficulty. In this section, participant responses to the open-ended questions of the post-task and task comparison questionnaires are referenced qualitatively to provide additional insight.

5.3.1 The relationship between task difficulty and perceived learning potential

Reasons proposed for the high difficulty of the VW HIGH task are somewhat as

predicted: relating to the unusually high demands of controlling an avatar, navigating the VW and manipulating specific items in the VW. For instance, comparing the VW HIGH task with the other two VW tasks, only the VW HIGH task required participants to place blocks and manipulate the environment. Interaction was limited to only navigating through the environment for the other two VW tasks.

As a concrete example of the difficulties that the VW HIGH task presented to participants, consider the following responses:

Participant 6 (VW HIGH comment): 操作方法は学んでなかったので難しかった。 [*It was difficult because we hadn't be taught how to control the characters in the game*.]

Even though there was an orientation class before participants engaged in the VW tasks, the class was not exhaustive and did not cover all of the intricacies of controlling their avatars in the VW. Therefore, the Room VW task was the most significant test of participants' knowledge of avatar and world manipulation, a point which Participant 6 brings up above.

Participant 7 (VW HIGH comment): 位置情報や方向を正しく伝えるのは母語でも大変。それが外国語で勉強できたのはとても有意義。 [*It is difficult to explain the position and direction of objects even in our native language. Doing this activity in English was very beneficial to our language skills*.]

Participant 7's comment here refers to the high cognitive demands of orienting oneself to the VW in order to give appropriate directions to an interlocutor, something that was required of participants in both the MID and HIGH complexity VW tasks. Participant 7 continues to state that the difficulty of the task was a contributing factor to her perception of its benefits for learning. This backs up the finding that was discovered in the statistical analysis of the post-task questionnaire: that the perceived-to-be highest complexity task was also considered to have the most learning potential.

Additional responses followed this pattern, stating that the VW HIGH task was more difficult than the FTF equivalent, but that despite this, the VW task may be more useful as a study tool (Participant 18), and more fun to complete (Participant 15):

Participant 18 (Room decoration comparison comment): オンラインの方が難

しかったので英語の勉強になった。 [*I think the VW version was more difficult, so it was better English practice for us.*]

Participant 15 (Room decoration comparison comment): マイクラの操作方法が分からなければ、オンライン活動の方が難しいと思うけど、操作ができれば紙よりは楽しい。 [*I think that the VW version would be challenging if you are not used to the controls in Minecraft. However, if you know the controls, I think it is more fun than the paper version*.]

However, results for the spoken data do not support participants' perceptions. Opposite to their perceptions, for these learners, their output was more fluent and complex when completing the FTF tasks. Why then do they feel that the VW tasks have a higher learning potential when their spoken output does not reflect this? Perhaps participants' perceptions of learning potential does not relate to their own output, but of the cognitive difficulty of the task only. As found in the statistical analysis of post-task questionnaire data, the cognitive demands of the VW tasks were perceived to be higher than the FTF tasks, statistically significantly so at the MID ($F(1, 19) = 18.10$, $p < .001$), and HIGH ($F(1, 19) = 29.10$, $p < .001$) complexity levels. There may be a simple, positive relationship between learners' perceptions of task difficulty and a task's learning potential where the more difficult a task is perceived, the more learning potential it is also considered to have.

5.3.2 Cognitive demands reduced by the immersive environment

The map used in the FTF MID task was considered difficult to read, and not as explicit as undertaking the task in the VW. In other words, the VW provided learners with a rich, 3D environment with a number of buildings that they could make reference to in order to help direct their partner. However, the FTF version only provided them with a 2D, top-down/birds-eye view map. The difference in clarity between the reference materials could explain why participants focused their attention on motor skills rather than language for the FTF task – because they were focused on parsing the impoverished 2D representation of the world (Figure 34).

Figure 34: The 2D map used in the Directions FTF task with possible destinations numbered 1 – 6.

A number of responses to the open questions provide backing to this point.

Participant 16 (FTF MID comment):

ビルなどを参考にできなかったので、オンライン活動よりは難しかった。 [*This was more difficult than the VW version because we didn't have buildings to use as a reference.]*

Participant 6 (FTF MID comment):

マップがちょっと見づらかった。 [*The map was a bit hard to see.*]

Participant 4 (MID task comparison comment):

マイクラでやった方が簡単だった。 [*It was easier to do this activity in Minecraft.*]

In further detail, reviewing responses to the open question of the MID task comparison questionnaire, 12 responses specifically mention the positive affordances of the VW (

Table 80). Among these responses, a number also mentioned that it was more fun and engaging to complete the task in the VW. Although the FTF task was predicted to be the simpler task based on task conditions, it may be argued that the virtual terrain did not place additional cognitive demands on participants, but instead provided semiotic resources in the form of easily distinguishable landmarks and objects for learners to refer to in their endeavour to guide an interlocutor during task performance. Participant 5, 14, and 19 mention this point explicitly in their comments.

Table 80: Participant responses to the open-ended comment section regarding a comparison of the VW and FTF version of the Directions task pair (English translation).

Participant Number	Comment
3	It was easier to do this activity in Minecraft.
5	We can see each other's expressions when we do it face to face, so this is better. In Minecraft, it is easier to look at things and speak about them.
6	It would be easier to do the paper task if the map was better.
7	I was able to give directions more easily with the virtual version, but the directions were very basic.
8	I think it can be better to use the 3D environment for learning directions
10	It was difficult to control our character in the VW version, but I think it was more fun.
11	I think the VW version was better because we can see our partners position easier
12	It was easier to do this in the VW than face to face because we can move around the map.
13	We were able to walk around the buildings in the virtual world version, so it was better.
14	Compared to the paper version, the virtual world had a lot of visual information to help us do this activity, so I felt like I spoke more English
15	It was difficult to control the character in Minecraft, but the environment was very immersive so much more fun than doing the paper version.
16	It was hard to give directions in Minecraft because our perspective was different (due to having two screens). I think it would be better to do this activity on only one computer.
18	I felt it was easier to do with Minecraft because it felt like a real city.
19	Doing this activity in Minecraft was better because we had real streets to direct our partner down.
20	I feel that it was better to do this activity whilst actually moving around a map. It felt much more authentic to do it in Minecraft.

Having explored the cognitive demands placed on learners for both modes of communication, I now turn my attention to the specific affective influences of the VW which were recorded in the post-task questionnaires. The following sections explore the concepts of willingness to communicate and motivation with a focus on how learning in immersive

3D environments may positively affect learner motivation.

5.4 AFFECTIVE AFFORDANCES OF LEARNING IN VIRTUAL ENVIRONMENTS

Motivation is considered an essential factor in the literature on learning with video games (e.g., Baltra, 1990; Boyle, Hainey, Connolly, Gray, Earp, Ott, & Pereira, 2016; deHaan, 2005). Suh, Kim and Kim (2010) go as far as to say that in their study, motivation was one of the most critical variables in determining learner performance. Additionally, findings suggest that motivation can be fostered through the use of digital games to teach languages (Anyaegbu, Ting, & Li, 2012; Voulgari, & Komis, 2011). However, Filsecker and Bündgens-Kosten (2012) caution researchers that motivation *to play* should not be confused with motivation *to learn* and that cognitive engagement with the subject matter is a more important goal than merely motivating learners to play (p. 64).

One item on the post-task questionnaire was designed to measure participants' enjoyment of tasks. The VW tasks had higher mean scores than the FTF tasks at all task complexity levels despite the increased cognitive difficulty posed by the modality. Additionally, the mean scores recorded for the item regarding task enjoyment on the post-task comparison questionnaires backed up this finding: of the two tasks in a task pair, the VW tasks were considered more enjoyable than the FTF equivalents at all complexity levels (Section 4.1.2 above). Further exploration revealed that qualitative data seemed to support this finding where responses to open-ended questions mentioned that the affordances of the VWs such as the opportunity to engage in authentic language use, the immersive nature of the environment, and ease of meaning-making due to recognizable symbols (such as the 3D terrain, non-player characters, or controllable avatars) had a positive effect on learners' willingness to engage with the materials and enjoyment of the VW tasks.

As a concrete example, in the open-ended question of the post-task comparison questionnaire asking participants to compare the two-spot the-difference tasks (Table 81), responses suggest that the task provided participants with more agency as they controlled their avatars and navigated the virtual terrain (Participant 10, 11, 15, 18 and 20). The affordances for exploration of the VW were also considered a positive by Participant 10 who writes, "*The FTF version was just one piece of paper so we can see everything in one glance. So, there was more to talk about with the VW version*." The immersive environment was thus

considered an important contributing factor to English development for this student. This finding echoes those by Liou (2012), where despite frustrations with Internet connection issues, her students felt that studying in Second Life was an authentic environment for communication and language development.

Table 81: Responses to the open-ended question section of the spot-the-difference task comparison questionnaire (English translation).

Participant Number	Comment
2	It was easier to see what the characters were doing in the online version.
5	I was relaxed because I was working with only one person. Other people couldn't hear me.
6	I thought I could concentrate more with the paper version.
7	When doing this activity in the virtual world, we cannot see our partner, so we have to concentrate on pronunciation much more.
8	It can be more difficult to use tools for language learning like games, but you get to experience things that you wouldn't normally be able to in reality, so that is a large benefit of using games.
9	I wasn't so nervous when I did the VW version.
10	The FTF version was just one piece of paper so we can see everything in one glance. So, there was more to talk about with the VW version.
11	We had to search for the characters in the VW version, so it took longer, but it was more fun.
12	I think the VW version was much harder because we had to control the character.
14	I felt more motivated to speak more English when doing the VW version, so it was better for me.
15	they were both quite fun, but it was more fun to do this activity in Minecraft
16	I think it is a good idea to do the VW version after the FTF version so that we can learn what to say first.
17	It was easy to see all the characters on one piece of paper and much harder to do this in Minecraft because we had to search for the characters.
18	It was fun to search for the characters in Minecraft
20	When doing the FTF version, we didn't have much to talk about other than reading a sentence. I felt like I could communicate more with the VW version.

In contrast, some participants, such as Participant 12 commented that "*I think the VW version was much harder because we had to control the character*," a comment that on first appearance suggests that the modality, or rather, the technical aspects of conducting the task

in the virtual environment produced additional, unnecessary cognitive demands. Indeed, looking at Participant 12's responses to the first part of the questionnaire, they also show a preference for the FTF version of the spot-the-difference task, as it was perceived to provide more opportunities to speak English, and for learning to occur (Table 82).

Table 82: Responses to the spot-the-difference task comparison questionnaire.

Participant Number	Pushed output	Task Difficulty	Language learning potential	Enjoyment
12	4	1	5	2

Note: 1 = virtual world preference, 5 = face-to-face preference

This example indicates that although in general the VW tasks seemed to promote participants to want to engage with the subject matter more, this is not true for all participants. Care must be taken when considering the use of such digital tools in the classroom, as the cognitive demands of the game may outweigh the learning potential for some students such as those inexperienced with digital games and virtual environments. One concrete example of this in the literature is Rama et al. (2011) where learners of Spanish were instructed to play the MMORPG *World of Warcraft* as an extracurricular activity, join a Spanish-speaking guild, and complete activities in the game world with Spanish speaking players. Their findings suggested that differences in participants' performance were caused not by a difference in Spanish language ability but prior knowledge of the game environment. Although one particular participant in the study had advanced Spanish language skills, she struggled to produce output as she had difficulties dealing with the complex mechanics and controls of the game world (p. 336). Additionally, Lee (2016) warns that although CMC tasks may be motivating for language learners, the cognitive demands of multimodality, technical difficulties, and lack of computer literacy may make some learners anxious and thus affect task performance (p. 83).

5.4.1 Willingness to communicate and the VW tasks

Three responses to the post-task comparison questionnaires were coded with VW-WTC, which was a code reserved for comments that specifically mentioned that a task motivated participants to communicate with their partners. These three responses are presented in Table 87 below:

Table 83: VW-WTC coded responses to the LOW complexity task comparison questionnaire (English translation)

Participant Number	Comment
5	I was relaxed because I was working with only one person. Other people couldn't hear me.
9	I wasn't so nervous when I did the VW version.
14	I felt more motivated to speak English when doing the VW version, so it was better for me.

Taking each of the comments in turn, Participant 5 indicates an increase in willingness to communicate due to the safe context provided by the VW. Unlike the FTF tasks, this participant finds the closed, anonymous-like environment relaxing, possibly resulting in a reduction of social anxiety. Participant 9 did not provide enough information to make any substantial claims as to why he was less nervous during the VW version of the spot-the-difference task, however it may be hypothesized that it was due to the anonymity afforded by the VW as found in Reinders and Wattana (2015).

Finally, Participant 14 writes explicitly that he felt more motivated to speak English when carrying out the VW task. Unfortunately, however, there is no additional background information as to why he wrote this. One hypothesis is that described above: the VW tasks provided a safe place for this participant to practise using English. As an alternative hypothesis, however, it could be related to the particularly strong, positive motivational affordances that have been associated with digital game play.

6 CONCLUSION

6.1 IMPORTANCE OF THE STUDY

The current study explored how learners' oral task performance differs when conducting interactional tasks in both face-to-face and virtual environments. This was identified as an important avenue for CALL research as there are numerous studies which explore the affective and cognitive benefits of using SCMC in language learning contexts, but few that focus explicitly on oral interaction (for a review, see Ziegler, 2016a). Additionally, as CALL has progressed from technology as tutor, to methodology or ecology and communication technologies have been widely accepted and integrated into educational realms of the industrialised world (Reinhardt & Thorne, 2016), it is critical that we understand how the medium may affect learner output compared to more traditional contexts such as FTF communication in classrooms.

It was argued that VW-based communication exists on the continuum of SCMC and has received little attention to date. Text- or chat-based SCMC has received the bulk of the field's attention (Warschauer, 1995; Iwasaki, & Oliver, 2003; C. Blake, 2009), followed by oral SCMC (Satar & Özdener, 2008; Yanguas, 2010; Yanguas; 2012; Yanguas; 2014). As late as 2017, there are papers appearing in the CALL literature which compare learner performance as they undertake text-based SCMC and oral FTF communication (Kim, 2017). Thus there is an imperative for exploring oral SCMC in more detail, particularly interactions that occur within complex VWs. To the best of my knowledge, there are still no studies that specifically compare the two modes of communication investigated here (FTF and VW-based SCMC). This area of research was therefore identified as being relatively unexplored in the literature on CALL where comparisons of SCMC and FTF communication generally focused on the use of text chat, and when oral communication was focused on, the tools used to mediate communication were VOIP software (such as Skype).

In this study, the approach taken was to analyse learner speech quantitatively in terms of complexity, accuracy and fluency using a number of appropriate measures. In order to determine how modality or task complexity affected learner output, three sets of task pairs were created to be LOW, MID and HIGH complexity based on the operationalisation of task conditions. Among the task pairs, the FTF version of a task was considered to be less

cognitively demanding than the VW equivalent. Task complexity predictions were validated by assessing learner *task difficulty* perceptions. This refers to learners' subjective perception of how cognitively demanding a task was, a construct that does not always match the instructor's prediction. In this study, the tasks that were designed to be more complex were considered more difficult, a finding that echoes those of Sasayama (2015, 2016).

Results of statistical analyses on the transcription data suggested that modality had little effect on learner output in terms of complexity or accuracy. Fluency, however, was statistically significantly affected by modality, where it was found that due to the affordances of the environment, learners were less fluent than when they completed the FTF tasks. Thus, one importance of the study is in showing that the use of VWs may not be the best option for instructors interested in promoting their learners' oral fluency, at least in relation to the types of tasks that were used here. In terms of task complexity, findings suggested that learners' familiarity with task content may have had more of an impact on output accuracy and complexity than modality. This was because the MID level complexity task promoted learners to produce statistically significantly more accurate output. Thus, findings highlight the importance of learners' prior experiences on determining task performance.

One unexpected result of this study is that learners tended to enjoy the VW tasks despite these tasks being considered more cognitively demanding and therefore more difficult than the FTF tasks. This finding helps answer RQ2 and implies that the relative ease of conducting simple FTF tasks may not be as motivating as those that allow learners to engage in tasks in an immersive virtual environment. Technology-mediated tasks may, therefore, help lower the Affective Filter and help learners engage with content, a point that has significant implications for instructors who teach learners lacking in intrinsic motivation to study a foreign language.

The conclusion chapter is structured in the following way: First I introduce the implications of this study for teachers that may be interested in exploring the use of virtual environments in their own contexts for improving their students speaking proficiency. After that, the limitations of the study are presented in detail. Limitations inform the rationale for future research and so are presented first. The dissertation concludes with future research considerations.

6.2 IMPLICATIONS FOR TEACHERS

The current study helped to expose the affordances of VW-based communication as part of a TBLT approach to SLA. However, it also uncovered the possible hindrances to successful language development that this modality poses; specifically: a decrease in output fluency and to a certain extent complexity. This section highlights the benefits and limitations of both modes of communication.

6.2.1 Considerations for the use of virtual worlds

The virtual world tasks used in this study were designed to allow learners to explore and manipulate a virtual environment, thus promoting agency and engagement with the content of the tasks. The controls and motor skills required to do this—use of a mouse and keyboard as well as a headset for communicating with interlocutors—may initially be a cause for concern for teachers, appearing to be too cognitively demanding of learners as they work towards task completion. However, as also found in Duquette and Hann, (2010) such technical difficulties are an invaluable source of meaning-focused communication as learners deal with communication breakdowns and problems provided by the environment. Indeed, in the case of this study, learners had to use the target language in exact ways in order to express correct meanings which affected fluency and overall language complexity. However, learners were pushed to engage in tasks for more extended periods of time, and think critically about their language use during task performance. Instructors may, therefore, explore such environments as a way of promoting purposeful language use as learners deal with problems as they arise.

Additionally, one significant finding of this study was that the VW tasks were considered more enjoyable than the FTF equivalents. For instructors, then, the immersive nature of such environments may help promote greater engagement with learning content for longer periods of time than is possible with FTF tasks. However, there is one substantial caveat to this point. Stockwell (2007) writes that teachers who wish to implement technology into their teaching contexts have a responsibility to be familiar with and discerning of any potential tools before they implement them. In other words, if teachers are not aware of how to use a particular technology, they should not attempt to use it with students. I also hold this position and have therefore endeavoured to become proficient in using technology for the purpose of conducting this research and broader teaching practices.

Activities that instructors must undertake to create VW tasks are considerable. For example:

- Learn how to navigate the environment,
- Discern what affordances the environment provides (what can and cannot be created within the VW),
- Learn how to design tasks that utilise those affordances (learn how to manipulate the VW),
- Make those tasks and relevant support materials,
- Spend class time teaching learners how to use the technology,
- Learn how to overcome technical difficulties.

Whether instructors are willing to invest the same effort, and even if they *should* invest such effort to learn about virtual environments is dependent on instructional aims and teaching context. As mentioned by Marklund and Taylor, "game-based learning processes are demanding on teachers, requiring them to take on many different roles, each of which requires a specific skillset" (2005, p. 367). They also note that integrating games into curricula can still be considered a laborious and complicated process.

Related to this point is the issue of continuing professional development in the field of language teaching, and in particular CALL. The issue is not new and has been written about by Garrett (2009) who laments the lack of support structure from institutions for teachers interested in exploring technology use in their teaching contexts. However, due to the normalisation and widespread adoption of the internet, there are a growing number of affinity spaces (Gee, 2007) available for teachers to become familiar with and participate in situated practice around technology for educational purposes. For instance, directly related to the present study and the software employed: *Minecraft*, Kuhn and Stevens (2017, p. 753-754) wrote about how they created a public community of practice composed of language teachers in order to solve the problem of how to "engage young language learners in the digital world they inhabit and in which video games figure largely."

Subsequently, I argue that instructors should consider their teaching environment before deciding whether it is worth investing time in developing learning content in a VW. For online courses such as those that cater to learners that cannot be co-located, VWs offer many of the same affordances of real-world, face-to-face classrooms. However, for

instructors whose L2 learners are physically co-located, and only communicate among themselves (as was the case in this study) the effort required to create, introduce, and conduct lessons or activities that utilise complex 3D environments may not produce learning gains that match the effort exerted.

However, it may be argued that although I created tasks that utilized the affordances of the online, networked virtual environment, I failed to utilize the largest affordance – the opportunity to connect learners to other people from outside the physical location. Therefore, VWs have the affordance for connecting monolingual learners to native speakers or other non-L1 speakers (Canto et al., 2014; Milton et al., 2012). However, in considering this option—to work collaboratively with participants from outside the instructional context—there may be even more demands placed on the instructor such as recruiting appropriate participants, applying for and gaining consent to conduct such classes from policymakers, and pre-empting and managing misunderstandings between cultures.

6.2.2 Affordances of face to face interaction

Research regarding the language learning affordances of various technological tools makes up a significant volume of the CALL literature. For instance, and in relation to the current study, Cornillie et al. (2012) conducted a meta-analysis to investigate how research interest in computer games for language learning had grown since the 1980s. They found a sharp rise in interest over the last decade (from 2001-2010). Additionally, of five research categories, the predominant avenue of research for this period was the design of language learning games and game-like environments followed by experimental, lab-based research. The pedagogical implications of using games was second lowest. The authors also stated that there are few empirical studies on game-related research in the CALL field. Additionally, scholars argue that there is a techno-utopian movement in the field of educational technology research, with a focus on uncovering the affordances of technology as a way of improving education (Tobias, Fletcher, Dai &Wind, 2011). Such is done without considering non-technological alternatives, or even appropriate pedagogical interventions around the use of technology (Selwyn, 2011).

The aim of this dissertation was to uncover the affordances of VWs for language learning and teaching in an instructed EFL context. However, this was not done in isolation

of non-technological alternatives. Tasks were designed for both traditional FTF and technology-mediated instruction, taking into account the specific possibilities and limitations of each mode. In this section, the affordances of the FTF mode for language learning is considered and briefly introduced based on a comparison of learner performances in both modalities.

Findings from this study suggest the FTF tasks push learners to greater fluency and can be completed in much shorter time periods than VW tasks designed to promote the use of the same linguistic structures. Focusing on the last point in particular, in this study, the face-to-face tasks were on average completed in half the time that it took to complete the VW equivalents (Table 84). With limited classroom time available, it may be argued that the positive affordances offered by virtual environments (such as positively affecting motivation) may not outweigh the time requirements of their implementation. Connected to this is that findings here highlight a need for further research on VW task design to help promote rather than hinder fluency.

Table 84: Completion time for all tasks with a comparison of FTF and VW completion times

(as a percentage).

Task	Mean time to complete (in seconds)
Spot FTF	468.50
Spot VW	903.50
FTF task completion time	52% of VW task
Room FTF	646.40
Room VW	2100.50
FTF task completion time	31% of VW task
Directions FTF	681.40
Directions VW	1066.10
FTF task completion time	64% of VW task
Average difference in completion time	FTF tasks took 49% less time to complete

For instructors, the positive affordance of face-to-face instruction (mediated by instructor-designed worksheets and materials) is that compared to VW tasks, they require less time to design and implement. This is not an insignificant benefit. With rapid prototyping approaches to task and lesson design, this allows instructors to design and implement a pedagogical intervention more rapidly, gather feedback and iterate their designs at a much faster rate than is possible when using the VW technology.

In summary, both modes of communication seem to offer unique affordances for promoting second language oral skills development. However, it is unclear whether one mode should take preference over the other, as their implementation depends on the context and curricular goals. Additionally, as reviewed in Skehan (2016), task conditions are not the only factor that may affect the possible learning potential of tasks, or task performance. Additional pedagogic activities sequenced alongside tasks may be more influential in determining task performance than task conditions themselves. In other words, tasks do not

exist in a vacuum but are embedded in a broader educational ecosystem. Concretely, examples include the *wraparound* activities (which can be considered the pre and post tasks of TBLT) instructors choose to implement (Sykes & Reinhardt, 2013), the instructor's personal philosophy and approach towards SLA pedagogy (TBLT, multiliteracies, SCT, etc.), and the overall social and cultural climate of the classroom (i.e. the motivational characteristics of learners, whether they have the same L1 or not, their ages, prior experiences with the L2, etc.).

6.3 LIMITATIONS

The limitations of the current study link directly to suggestions for further research in the field. As such, limitations appear here, before making suggestions to other researchers regarding possible future research directions. I offer five significant limitations: 1) task design, 2) CAF measures employed in this study, 3) post-task questionnaire design, and 4) sample size, 5) and participant proficiency levels. These are introduced in order.

6.3.1 Task design

Although tasks in a task pair were designed to be as similar as possible in terms of their cognitive demands, it was difficult to predict how learners would engage with them, even after conducting a pilot study. The first, key question, however, is "Were tasks in a task pair comparably similar?"

Both tasks in a task pair required learners to navigate task goals verbally, where task goals were equal for both versions of the same tasks. Individual learner objectives were also split equally, for example, the inclusion of three destinations to navigate to for each participant for both online and offline tasks. This was done as a way to promote an equivalent volume of output for each participant in a pair. Thus, on first glance, tasks appear to be similar, at least superficially. However, a number of factors undermine this initial similarity: 1) learners interpretation of task goals or "task as process," 2) unconsidered, emergent difficulties of conducting tasks in the virtual environment, 3) inevitable differences due to the affordances of the media. I will now discuss each of these points in turn.

6.3.1.1 Learner interpretations of tasks

As Breen (1989) originally defined, tasks can be viewed as either a designed

pedagogical artefact ("task-as-workplan") or as something which learners interpret and reconstruct as they carry out the task in real-time ("task-as-process"). The mismatch between the two interpretations can make studying task performance between different tasks problematic. For instance, although an instructor may have designed two tasks to be interpreted and carried out the same way, learners may not have the same interpretation for each task, resulting in individually different task performances or even outcome alignment that differ largely from the instructors/designers intended outcome.

As suggested then, individual learners may interpret tasks differently resulting in different performances between participants or pairs of participants as they complete the same tasks. It is therefore near impossible to claim that any two tasks will promote precisely the same task performance for all learners. The phenomenon is well known in the field, where the areas of performance that are prioritized are out of the hands of the task designer, as learners themselves focus on which areas of performance they should prioritize (Harris, 2005). This is a limitation of the current study, and indeed any study that attempts to compare learner performances as they carry out tasks designed to be different in ways that can be measured.

6.3.1.2 Unforeseen complexities afforded by the VW

In the case of the current study, a factor related to the complexity of the virtual world emerged which seemed to affect task complexity. This is partly my fault as the designer of the online tasks. My own ability at manipulating the VW used in this study clouded my judgement of how hard learners would find controlling their characters in the environment. My "curse of knowledge" (a term coined by Camerer, Loewenstein, & Weber, 1989) regarding the game world created a mismatch between my assumptions of task complexity and learners' ability. This manifested itself most obviously with the VW HIGH task which was considered a lot more complex than the FTF equivalent, in ways that were not originally designed. As described in Section 5.2.1.1, manipulating the 3D space, such as the in-game blocks was much more complex than originally anticipated. In order to keep tasks as similar as possible then, more care should have been taken to ensure that the number of elements that learners manipulated in each task were equal.

6.3.1.3 Designing tasks to make use of the environments' affordances

All tasks used in this study were created to make the best use of the affordances of the medium, which in hindsight means that the two tasks in a task pair may be considered very different. This was a specific design consideration and one that I stand by. As mentioned in the introduction to this paper (Section 1.2), the literature on game-based language learning often has learners conduct tasks that do not fully utilise the affordances of the environment they are conducted in (for a review see Swier, 2014). This was considered as both a challenge to design tasks that make use of VW affordances and also as the impetus for answering the research questions of this paper. That is: what is the difference in learner output when the complete tasks that are completed 1) face-to-face and 2) in a virtual environment where tasks have been *specifically designed to utilize affordance of that environment*.

As a specific example taken from this study, the VW MID (directions) task required learners to navigate the virtual terrain as opposed to referring to a 2D map of the same town (as done in the FTF equivalent). If I had instead designed the VW task for learners to merely make reference to a 3D map displayed on their screen, the argument could be made that there is no need for that task to be carried out in the VW; VOIP communication would suffice. Consequently, I would not be comparing the two modes of communication I initially set out to compare (FTF versus VW-based task performance). Instead, I would have been comparing FTF and VoIP based SCMC performance. Thus, and in summary, because of the specific design considerations employed in this study (that tasks would be designed to make affordances of the medium there are conducted in) differences in task complexity were somewhat unavoidable.

Related to this point is that the task procedure, or way in which each lesson was conducted, may have also affected the reliability of results. Whilst care was taken to provide learners with a similar lesson procedure each week; the pre-task activities did differ. Whether this was an influence in determining learners' performance is unknown. Regardless, care should have been taken to ensure that pre-task activities were equivalent for all tasks.

Finally, the immersive nature of the VW tasks appeared to be an indicator of motivation to learn via this modality. However, it should be noted that the level of immersion of these tasks could be improved. As mentioned in the section on task design (Section 3.3.1), the spot-the-difference and directions VW tasks required participants to make reference to

worksheets outside of the virtual environment during task performance, potentially reducing the level of immersion for these tasks. Reference to external worksheets was included due to practical reasons: the VW used in this study features no way of providing learners with images or text files to make reference to during task performance. As a concrete example, it was impossible to provide learners with a map of the area they were traversing in the VW MID task, thus the need to provide a map externally. The widespread adoption of virtual reality (VR) and head-mounted display systems may offer more robust methods of providing task-plans to participants within the domain itself, i.e. external worksheets may be administered within the VR environment for participants to complete during task performance (if needed at all). Additionally, modern VR offers even more immersive experiences than the modality used here; experiences which are highly engaging, produce strong emotional reactions and a feeling of presence (Wilcox, Allison, Elfassy & Grelik, 2006; Sykes, Oskoz, & Thorne, 2008). VR may, therefore, provide further opportunities for increased levels of immersion which may consequentially lead to improved motivation to engage in presented subject materials.

6.3.2 CAF measures

CAF measures used in this study were selected based on their appropriateness after reviewing the literature of similar studies. However, there are of course a plethora of other measures that I could have employed. Using different measures may have provided additional and/or different insights into the relationship between task complexity and output performance. For instance, fluency was only measured in terms of temporal fluency. Whilst words-per-minute is a theoretically sound measure of fluency; there are also other measures of fluency that could have been included in this study such as the *vocal* fluency measures (number of false starts, reformulations, repetitions, etc.). Additionally, as output fluency was the only measure to be significantly affected by modality (complexity and accuracy showing no significant differences), there is a need for further exploration of how modality effects learner fluency in greater detail.

6.3.3 Post-task questionnaire design

A number of post-task questionnaires were employed in this study, where learners were required to compare the two tasks in a task pair. Comparisons were made with a number of measures such as the perceived difficulty, enjoyment, and perceived learning gains of both

tasks. However, on further reflection, it may have been helpful to provide learners with a place on the post-task questionnaires to explicitly rank all tasks in order of difficulty. This may have provided additional data to rank the tasks in order of perceived difficulty.

Furthermore, the third question of the post-task questionnaire was designed to promote learners to evaluate task design elements, providing insights to the particular affordances or difficulties related to the mode of communication. However, due to the vague wording of the question learners interpreted "improvements" as relating to how their own, personal performance may be improved. Thus, as is common in studies which employ questionnaires, it should be recognised that responses to this question also featured issues that learners had with language, task, and environment, which may be considered a minor limitation.

Additionally, and again, concretely, there was an issue with the post-task question regarding learners focus. This question was adapted from the Cognitive Load Subjective Experience Questionnaire developed by Paas, Van Merrienboer, and Adam, (1994), however, my interpretation of learner responses is that the question was not explicitly worded enough, and thus results should be rejected. In hindsight, this question should have been separated into two similar questions:

- How much effort did you put into thinking about English during this task?
- How much effort did you put into
    - Controlling you avatar or manipulating the virtual landscape? (VW tasks)
    - Manipulating the physical objects associated with this task? (FTF tasks)

The final issue regarding the post-task questionnaires is the level at which these learners are able to reflect on their own learning. Whilst it is assumed the learners provided honest answers in the questionnaire, there is a lack of depth, which again complicates the formation of firm conclusions regarding the connection between task complexity, task difficulty, and spoken performance. Related to the lack of depth afforded by a written questionnaire, this study may have also benefitted from gathering data via semi-informal, semi-structured interviews with participants in order to triangulate data further and improve the internal validity of findings.

6.3.4 Number of participants

After having to remove the data collected from a number of pairs due to absences, and defects with recordings, the current study had a sample size of only 20 participants. Additionally, the participants were not selected by the researcher but were members of intact classes who had chosen to take the elective course. As with any study that relies on statistical analyses, low sample sizes can be a cause for concern when analyzing data and making generalizable claims of findings. In order to advance our knowledge on the relationship between modes of communication and learners' oral task performance, particularly how tasks designed to be carried out in complex virtual environments may affect learner output, further studies with more substantial sample sizes are needed.

6.3.5 Participant proficiency level

The participants of this study were non-English major university students in Japan. Although no standardized English test scores were used to establish their English proficiency (such as TOEIC, TOEFL or IELTS), it is assumed that the majority of them are beginner level learners. This in itself is one limitation of the study: that there was no thorough investigation of learners' English proficiency before the intervention. Data was collected but was based largely on measures on a questionnaire that asked participants to provide subjective evidence of their own proficiency. Apart from only two learners (Participant 8 and 20 have taken the TOEIC test), none of them had taken a proficiency test and had had minimal contact with English outside of the compulsory classes they took at the secondary level of their education. In summary, although no standardized tests were used to assess learners' English proficiency, preliminary data was collected via a self-evaluation questionnaire where it was discovered that they were mostly homogeneous in their past experiences with English.

Results of the current study, then, may only be tentatively generalized as indicating how beginner learners' performance may differ when conducting FTF and VW-based tasks. Looking specifically at the difference in performance between proficiency groups is a much-needed extension of the current study.

6.4 IMPLICATIONS FOR RESEARCHERS

The limitations section of this study helps inform future researchers of possible ways in which it could be expanded on. The first is in selecting an appropriate virtual environment.

There are a plethora of tools one may choose from, and therefore the researcher must conduct a thorough investigation in order to determine which platform/software is appropriate for the research they wish to conduct. As an additional requirement of researchers in this field, it should go without saying that it is paramount that they have an in-depth knowledge of any technology implemented in their study. There are a number of implications to not having such knowledge: 1) Researchers will not be able to effectively design tasks for an environment without having explicit knowledge of its affordances, 2) it will be difficult to gauge the level of task complexity for any tasks designed for the environment, 3) it would be difficult to identify how the VW caused problems for learners and affected their output.

Secondly, researchers should be prepared for unexpected factors to influence the level of cognitive demand presented by tasks. As seen in this study, there were a number of factors which seemed to affect cognitive demand including lack of understanding regarding the complexities of controlling an avatar, the time requirements of actions in the VW, and the (lack of) clarity of input materials (for instance the Directions FTF map). Therefore, there is a necessity of conducting post-task perception questionnaires in order to consider learners perceptions of task difficulty. In this study task complexity predictions and task difficulty perceptions coincided but that may not always be the case.

6.5 FUTURE RESEARCH

The findings of this study may be considered a starting point for further research into the field of technology-mediated TBLT with a focus on task design for promoting oral communication. The current study employed university-level, Japanese speakers of English with low-level proficiency as participants, and as such, the question remains whether replication of the study in other contexts would yield the same results. Additionally, findings here, although not universally generalizable, departed from the predictions of Robinson's CH, where more complex tasks seemed to hinder learners' output complexity and accuracy. One reason for this could relate to the low proficiency level of the learners. Indeed, the clauses per AS-Unit and words per AS-Unit measures of oral complexity had to be rejected because it was found that participants rarely produced utterances of more than one clause in length. With participants of a higher proficiency level, results may more closely match predictions. However, if such is not found, the applicability of the CH to VW-based

communication may require reconsideration.

Following, the task pairs designed for this study were markedly different from one another. For example, the spot-the-difference task was a typical two-way information gap activity, whilst the room decoration task was a one-way information gap, repeated for each participant in a dyad. This study may therefore be replicated with more closely related tasks, achievable by manipulating only a single task variable such as "number of elements." Additionally, it may be worth conducting research that only focuses on VW communication. Such studies may provide more detailed answers to the relationship between task complexity and learner output when conducting tasks in virtual environments.

An avenue for further research also exists regarding the finding that learners enjoyed the more mentally demanding VW tasks over the FTF tasks. Specifically, the following question raised by this study may warrant further investigation in order to improve our understanding of the motivational effects of VWs on language learners' willingness to communicate in the L2 and engage with subject materials:

What elements of the VW tasks did learners find motivating?

The effect that learning with technology has on learner motivation has been explored in numerous studies. A meta-analysis by Warschauer (1996) includes the following factors of technology as potentially motivating for language learners: novelty, individualized instruction, opportunities to express learner agency, opportunities for rapid, individualized feedback. Additionally, and in keeping with the current study, the use of complex virtual environments has been shown to positively affect language learners' motivation due to the opportunities to interact with native speakers (Thorne, 2008), the immersive nature of the environment (Choi, 2006), opportunities to receive feedback from peers (Rankin, Gold & Gooch, 2006), the low-stakes and fun nature of the environment (Reinders & Wattana, 2011), the lowering of foreign language anxiety (Wehner, Gump, & Downey, 2011) and the promotion of curiosity among learners (Malone, 1981).

Responses to the post-task questionnaires in this study did not seem to suggest that novelty was a contributing factor, but this point was not explored in any depth. Thus, further research is sorely needed to deepen our understanding of what elements of the VW increase learners' motivation to engage with tasks and their willingness to communicate.

Related to this question is "*How much influence did task design have on learners' enjoyment?"* For instance, in this study, tasks were designed to make use of the affordances of the environment, but if this were not the case, and tasks were designed to more closely resemble the FTF tasks, would they be as enjoyable? Additionally, as mentioned in Section 6.3.1.3 on the limitations of this study, the level of immersion that these tasks offered participants pales in comparison to those experiences afforded by modern VR systems. An empirical, comparative investigation comparing monitor-based VW and head-mounted display-based VR modes of communication may reveal that due to its similarity to the FTF mode, VR may offer learning potential that takes the best of both FTF and VW modes. In other words, learner task performance may match that of the FTF mode, but have the added benefits of 1) being able to conduct immersive, emotional experiences that would be impossible to replicate in traditional classrooms, and 2) increased enjoyment and motivation to engage.

Finally, future studies should investigate the relationship between modality and learning gains. The current study focused on learners' task performance, where I hope to have shown the language learning affordances of such domains and how teachers may want to approach task design in order to capitalize on these affordances. As part of this study, I also asked learners for their perceptions regarding the learning gains of the tasks and modes of communication. One significant finding is that learners were more enthusiastic and motivated to complete the tasks in the VW in spite of the fact that they were more cognitively demanding. Logically then, the assumption could be made that VWs offer the opportunity for increased learning gains as learners endeavour for longer at more difficult tasks. However, concrete learning gains were not accurately assessed in the current study. More rigorous research may help ascertain 1) whether there are differences in the potential learning gains between the various modes of communication and 2) if there are pronounced differences, how they may be utilized as a part of curriculum development and more precisely how they inform best practices for task design for virtual environments.

APPENDICES

APPENDIX 1 CONSENT FORM OUTLINE

Research aims

This class is a part of Mr. York's research on the use of virtual worlds as a teaching tool in English language education. Results will be submitted to the University of Leicester for the award of a PhD in Education. The research involves an analysis of your speaking skills when undertaking tasks in a virtual world. The broad aim is to see what effect virtual worlds have on your speaking skills when compared to similar tasks undertaken face-to-face. The data will thus be used to:

1. Assess the quality and quantity of your speech during tasks
2. Understand your opinion regarding task difficulty

Data collection methods

Your participation in this research will involve a variety of data collection methods. Data will be gathered in the form of audio recordings when you do tasks, both in the virtual world and face to face. All participants that give consent will be recorded. Questionnaires will also be employed to gather your opinions on all tasks. Although it is necessary to collect your student numbers throughout the data collection process (to match questionnaire results with audio recordings), no real names will be used in the final research thesis, only pseudonyms if at all necessary. Your anonymity will be kept.

Consent areas

- Recording of spoken audio during tasks
- Collect your opinions via several questionnaires
- To participate voluntarily
- You have the right to end your participation at any time during the semester without the need to give a specific reason.

Data handling procedure

I promise to do the following things regarding the handling of any data collected:

- I will store all data on a password-locked computer in my secure office space.
- I will respect your anonymity and not used any real names in the final report.
- I will give you the chance to view the results of any research before it is made public
- I will use collected data only for the purpose of this project.
- I will not use any collected data as a means to assess you on this course. In other words, your participation or non-participation in this research has no bearing on your final grade.

Participant's name_____ Date _____

Researcher's name_____ Date _____

APPENDIX 2: CODED RESPONSES TO THE OPEN-ENDED QUESTIONS OF THE INDIVIDUAL POST-TASK QUESTIONNAIRES BY MODALITY

SPOT-THE-DIFFERENCE TASKS								
Task	FTF	VW	language	FTF	VW	Environment	FTF	VW
fun	6	7	authentic	2	1	easier	2	2
simple	10	2	simple	7	5	WTC	2	1
difficult	0	0	difficult	4	6	technical	6	3
			communication	19	14	affordances	4	3
			pronunciation	1	0	fun	2	5
			words	6	1			
			grammar	5	2			
			phrases	2	0			
			translation	0	1			
			listening	4	0			
			non-verbal	3	0			
			JP	0	2			
			ZPD	0	2			
ROOM DECORATION TASKS								
Task	FTF	VW	language	FTF	VW	Environment	FTF	VW
fun	5	4	authentic	3	3	easier	1	0
simple	4	0	simple	0	1	WTC	0	0
difficult	2	8	difficult	7	18	technical	2	16
			communication	5	15	affordances	4	5
			pronunciation	0	1	fun	1	2
			words	6	7			
			grammar	18	5			
			phrases	0	0			
			translation	0	0			
			listening	0	0			
			non-verbal	3	0			
			JP	0	0			
			ZPD	0	1			
DIRECTIONS TASKS								
Task	FTF	VW	language	FTF	VW	Environment	FTF	VW
fun	1	2	authentic	0	6	easier	2	0
simple	14	5	simple	7	8	WTC	0	1
difficult	2	2	difficult	1	5	technical	4	8
			communication	3	7	affordances	4	4
			pronunciation	1	0	fun	0	2
			words	8	3			
			grammar	9	1			
			phrases	2	6			
			translation	0	0			
			listening	0	0			
			non-verbal	0	0			
			JP	0	0			
			ZPD	0	0			

APPENDIX 3: CODED RESPONSES TO THE POST-TASK COMPARISON QUESTIONNAIRES FOR EACH TASK-PAIR

Sub-category	Spot FTF	Spot VW	Room FTF	Room VW	Dir FTF	Dir VW
cognitive-low	5	1	2	0	1	1
cognitive-high	0	5	0	8	0	2
WTC	0	3	0	0	0	0
technical	0	2	0	5	0	3
affordances	0	8	2	6	1	12
positive	1	9	0	6	1	10

APPENDIX 4: EXAMPLE LESSON WORKSHEET

Pre-task

1: Vocabulary lists

Think about the items that you have in each of these rooms. Try to make a list of 10 for each.

Living room

Bedroom

Kitchen

Bathroom

2: Videos

Please watch these videos of native English speakers doing the same task as you will do.

https://www.youtube.com/watch?v=zYoahJSkewg

https://www.youtube.com/watch?v=m51-FGwqOts

Make notes on what words they use

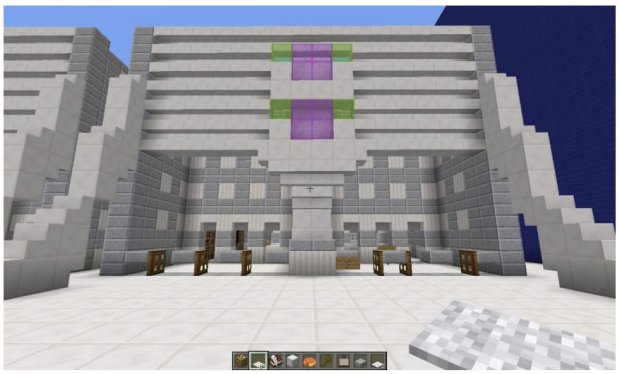
This will help you when you do the activity yourself. Please try and write:

- Minecraft block words
- Preposition words
- Phrases and other important vocabulary

Task

With your partner, please go to **/warp seminar**.

From here you can find the buildings to do this activity. One player must enter each side of the building:



Post-task & Homework (Report)

Recipe

We often see this grammar used in recipes. Here is an example. Please complete the cooking instructions with words from the table.

Add	Grate	Heat	Increase
Leave	Pour	Remove	Season

Preheat your oven to 240°C. _____ a some olive oil in a large frying pan over a medium heat.
_____ your onion and garlic and cook for about 10 minutes until soft.
Add the mushrooms with the leaves from a few of your thyme sprigs.
_____ the heat to high and fry for 5 to 10 minutes until the mushrooms go slightly crispy.
_____ from the heat, _____ in the zest of the lemon and _____ well.
_____ into a bowl and _____ to one side to cool.

APPENDIX 5: POST-TASK QUESTIONNAIRE (FTF VERSION) (JAPANESE VERSION)

FTF Questionnaire TASK _____ 学籍番号：_____

1. 今日の活動にはどれくらい精神的な努力をしましたか。

全然精神的な努力をしなかった。	あまり精神的な努力をしなかった。	どちらも思わない。	結構精神的な努力をした。	かなり精神的な努力をした。

2. 今日の紙ベースで活動するのは簡単か、難しいですか。

かなり簡単だった。	ちょっと簡単だった。	どちらも思わない。	ちょっと難しい。	かなり難しかった

3. 出てくる言葉はわかりやすいか。

かなりわかりやすかった。	ちょっとわかりやすかった。	どちらも思わない。	ちょっとわかりにくかった。	かなりわかりにくかった。

4. この活動で英語を話すことか、ものを書くこと、どちらに集中しましたか。

かなり書くことに集中した。	ちょっと書くことに集中した。	どちらも思わない。	ちょっと英語を話すことの方に集中した。	かなり英語を話すことの方に集中した。

5. 今日の活動は楽しかったか。

全くそう思わない	そう思わない	どちらも思わない。	そう思う	全くその通り

今日の活動で何かを学びましたか。学んだ事があったら書いてください

_____

_____

_____

_____

今日の活動で良かったところを書いてください

_____

_____

_____

_____

今日の活動で直した方がよいところがあったら書いてください

_____

_____

_____

_____

APPENDIX 6: POST-TASK QUESTIONNAIRE (VW VERSION) (JAPANESE VERSION)

VW Questionnaire　TASK _____　**学籍番号：** _____

1. 今日の活動にはどれくらい精神的な努力をしましたか。

全然精神的な努力をしなかった。	あまり精神的な努力をしなかった。	どちらも思わない。	結構精神的な努力をした。	かなり精神的な努力をした。

2. このバーチャルワールドで活動は簡単か、難しいか。

かなり簡単だった。	ちょっと簡単だった。	どちらも思わない。	ちょっと難しい。	かなり難しかった

3. バーチャルワールドで出てくる言葉はわかりやすいか。

かなりわかりやすかった。	ちょっとわかりやすかった。	どちらも思わない。	ちょっとわかりにくかった。	かなりわかりにくかった。

4. この活動で英語を話すことか、ゲームの操作か、どちらに集中しましたか。

かなりゲームの操作の方に集中した。	ちょっとゲームの操作の方に集中した。	どちらも思わない。	ちょっと英語を話すことの方に集中した。	かなり英語を話すことの方に集中した。

5. 今日の活動は楽しかったか。

全くそう思わない	そう思わない	どちらも思わない。	そう思う	全くその通り

今日の活動で何かを学びましたか。学んだ事があったら書いてください

_____

_____

_____

_____

今日の活動で良かったところを書いてください

_____

_____

_____

_____

今日の活動で直した方がよいところがあったら書いてください

_____

_____

_____

_____

APPENDIX 7: POST-TASK QUESTIONNAIRE (VW VERSION) (ENGLISH TRANSLATED VERSION)

VW Questionnaire **TASK** _____ **Student No.** _____

1. How much mental effort did you put in to complete this task?

None at all	Not much	Not sure	A little	A lot

2. How easy or difficult was the activity?

Very easy	A little easy	Neither	A little difficult	Very difficult

3. How easy or difficult was the video game's language to understand?

Very easy	A little easy	Neither	A little difficult	Very difficult

4. In this activity, did you concentrate more on speaking English or the game controls?

I focused a lot more on the game controls	I focused a little more on the game controls	Neither	I focused a little more on speaking English	I focused a lot more on speaking English

5. Was today's activity fun?

Not at all	Not really	Not sure	Yes, a little	Yes, a lot

What did you learn in today's activity?

_____

_____

_____

_____

What were the good points of this activity?

_____

_____

_____

_____

What do you think should be improved with this activity? If you have any suggestions, please write them below.

_____

_____

_____

_____

APPENDIX 8: POST-TASK QUESTIONNAIRE – FTF AND VW COMPARISON (ENGLISH TRANSLATED VERSION)

FTF vs VW Task comparison **Student No. _____**

Comparison of virtual world and face-to-face activites

Please put a circle in the box that best matches your experience. If there was no difference between the virtual world and face-to-face tasks, please put a circle in the centre "Neither" box.

	The virtual world version	A little more when doing the virtual world version	Neither	A little more when doing the face-to-face version	The face-to-face version
Spoke English					
Was fun					
Was difficult					
Was useful for learning English					

If you have any comments regarding the difference between the two tasks, please write it here:

_____

_____

_____

REFERENCES

Abrams, Z. L. (2003). The effects of synchronous and asynchronous CMC on oral performance. *Modern Language Journal*, *87*(2), pp. 157–167.

Ågren, M., Granfeldt, J., & Schlyter, S. (2012). The growth of complexity and accuracy in L2 French: Past observations and recent applications of developmental stages. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Investigating complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, pp. 143–169.

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: does it transfer to a new task? *The Language Learning Journal*, *39*(3), pp.

269-280, DOI: 10.1080/09571736.2010.545239

Alastuey, M. C. B. (2010) Synchronous-voice computer-mediated communication: Effects on pronunciation. *CALICO Journal*, *28*(1), pp. 1–20.

Allen, L. K., Crossley, S. A., Snow, E. L., & McNamara, D. S., (2014). L2 writing practice: Game enjoyment as a key to engagement. *Language Learning & Technology, 18*(2), 124–150.

Amiryousefi, M. (2016). The differential effects of two types of task repetition on the complexity, accuracy, and fluency in computer-mediated L2 written production: a focus on computer anxiety. *Computer Assisted Language Learning*, *29*(5), pp. 1050–1066. http://doi.org/10.1080/09588221.2016.1170040

Anyaegbu, R., Ting, W., & Li, Y. (2012). Serious game motivation in an EFL classroom in Chinese primary school. *TOJET: The Turkish Online Journal of Educational Technology*, *11*(1), pp. 154–164.

Appel, G., & Lantolf, J. P. (1994). Speaking as mediation: A study of L1 and L2 text recall tasks. *The Modern Language Journal*, *78*(4), pp. 437-452.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baierschmidt, J. (2013). A principled approach to utilizing digital games in the language learning classroom. *JALTCALL Journal*, *9*(3), pp. 307–315.

Balacheff, N., Ludvigsen, S., de Jong, T., Lazonder, A., & Barnes, S. (Eds.). (2009). *Technology-Enhanced Learning*. Springer Netherlands. https://doi.org/10.1007/978-1-4020-9827-7

Baltra, A. (1990). Language learning through computer adventure games. *Simulation & Gaming*, *21*(4), pp. 445-452.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development*, *53*(1), 86–107. https://doi.org/10.1007/BF02504859

Baralt, M. (2010) *Task complexity, the Cognition Hypothesis, and interaction in CMC and FTF environments.* Unpublished doctoral dissertation. Georgetown University, Washington, DC.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition*, *35*(4), 689–725. https://doi.org/10.1017/S0272263113000429

Baralt, M., & Gurzynski-Weiss, L. (2011). Comparing learners' state anxiety during task-

based interaction in computer-mediated and face-to-face communication. *Language Teaching Research*, *15*(2), pp. 201–229.

Bartle, R.A. (2003). *Designing virtual worlds*. Boston, MA: New Riders.

Bax, S. (2003). CALL - Past, present and future. *System*, *31*(1), pp. 13–28. http://doi.org/10.1016/S0346-251X(02)00071-4

Bax, S. (2011). Normalisation revisited: The effective use of technology in language education. *International Journal of Computer-Assisted Language Learning and Teaching*, *1*(2), pp. 1–15.

Beauvois, M. H. (1997). Computer-mediated communication (CMC): Technology for improving speaking and writing. In M. D. Bush & R. M. Terry (Eds.), *Technology-enhanced language learning*. Lincolnwood, IL: National Textbook Company, pp. 165–184.

Berger, P. & Trexler, S. (2010). *Choosing Web 2.0 tools for learning and teaching in a digital world*. Santa Barbara, C. A.: Greenwood Press.

Bialystok, E. (1994). Analysis and control in the development of second language proficiency. *Studies in Second Language Acquisition, 16*(2), pp. 157–168.

Black, R. (2005). Access and affiliation: The literacy and composition practices of English-language learners in an online fanfiction community. *Journal of Adolescent & Adult Literacy, 49*(2), pp. 118-128.

Black, R. (2006). Language, culture, and identity in online fanfiction. *E-Learning, 3*(2), pp. 170-184.

Black, R. (2009a). English-language learners, fan communities, and 21st-century skills. *Journal of Adolescent & Adult Literacy, 52*(8), pp. 688-697.

Black, R. (2009b). Online fan fiction, global identities, and imagination. *Research in the Teaching of English, 43*(4), pp. 397-425.

Blake, C. (2009). Potential of text-based internet chats for improving oral fluency in a second language. *The Modern Language Journal*, *93*(2), pp. 227-240.

Blake, R. (2000). Computer-mediated communication: A window on L2 Spanish interlanguage. *Language Learning and Technology*, *4*(1), pp. 120–136.

Blake, R. (2016a). Technology and the four skills. *Language Learning & Technology*, *20*(2), pp. 129–142.

Blizzard Entertainment. (2004). *World of Warcraft*. [PC video game]. Irvine, CA: Blizzard Entertainment.

Blume, C. (2019). Playing By Their Rules: Why Issues of Capital (Should) Influence Digital Game-Based Language Learning in Schools. *CALICO Journal*, *36*(1), 19–38. https://doi.org/10.1558/cj.35099

Blyth, C. (2018). Immersive technologies and language learning. *Foreign Language Annals*, *51*(1), 225–232. https://doi.org/10.1111/flan.12327

Bogdan, R. C., & Biklen, S. K. (1992). *Qualitative research: An introduction to theory and methods*. Needham Height: Allyn & Bacon.

Böhlke, O. (2003) A comparison of student participation levels by group size and language stages during chatroom and face-to-face discussions in German. *CALICO Journal*, 21(1): 67–87.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., & Pereira, J. (2016), An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education*, 94, pp. 178–192.

Breen, M. (1987). Learner contributions to task design. In C. Candlin and D. Murphy (Eds.), *Language Learning Tasks*. Englewood Cliffs NJ: Prentice-Hall.

Breen, M. (1989) The evaluation cycle for language learning tasks. In R.K. Johnson (ed.), *The second language curriculum*. Cambridge: Cambridge University Press.

Brown, J. S., Collins, A. and Duguid, P. (1989) Situated cognition and the culture of learning. *American educational research association, 18*(1): 32–42.

Bruckman, A. (1999). Can educational be fun? *Game developers conference*, 99, pp. 75-79.

Brumfit, C. (1984). *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy*. Cambridge: Cambridge University Press.

Bryman, A. (2015). *Social research methods*. Oxford: Oxford University Press.

Buendgens-Kosten, J. (2013). Authenticity in CALL: Three domains of "realness." *ReCALL*, *25*(2), 272–285. https://doi.org/10.1017/S0958344013000037

Burrows, C. (2008). Socio-cultural barriers facing TBL in Japan. *The Language Teacher*, *32*(8), pp. 15–19.

Bygate, M., (1996). Effects of task repetition: appraising the developing language of learners' in D.Willis and J.Willis (Eds.) *Challenge and change in language teaching*. London: Heinemann, pp. 136-146.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic*

tasks: Second language learning, teaching and testing. Harlow: Longman, pp. 23–48.

Bygate, M., & Samuda, V. (2005). Integrative planning through the use of task-repetition. In R. Ellis (Ed.), *Planning and task performance in a second language*. Amsterdam: John Benjamins Publishing Company, pp. 37-74.

Bygate, M., Norris, J. & Van den Branden, K. (2009). Understanding TBLT at the interface of research and pedagogy. In Van den Branden, K., Bygate, M. & Norris, J. (Eds.), *Task- based Language Teaching: A reader*. Amsterdam: John Benjamins, pp. 495-500.

Cadierno, T. (2004). Expressing motion events in a second language: A cognitive typological perspective. *Cognitive linguistics, second language acquisition, and foreign language teaching*, pp. 13-49.

Cadierno, T., & Robinson, P. (2009). Language typology, task complexity and the development of L2 lexicalization patterns for describing motion events. *Annual Review of Cognitive Linguistics*, *7*(1), pp. 245-276.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of political Economy*, *97*(5), pp. 1232-1254.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), pp. 1–47.

Candlin, C. N. (1987). Towards task-based language learning. In C. N. Candlin & D. Murphy (Eds.), *Lancaster Practical Papers in English Language Education: Vol. 7. Language learning tasks*. Englewood Cliffs, NJ: Prentice Hall, pp. 5-22.

Canto, S., de Graaff, R., & Jauregi, K. (2014). Collaborative tasks for negotiation of intercultural meaning in virtual worlds and video-web communication. In M. Gonzalez-Lloret, & L. Ortega (Eds.), *Technology and tasks: Exploring technology-mediated TBLT*. Washington, DC: Georgetown University Press, pp. 183-212.

Cerezo, L., Baralt, M., Suh, B.-R., & Leow, R. P. (2013). Does the medium really matter in L2 development? The validity of CALL research designs. *Computer Assisted Language Learning*, *27*(4), pp. 294–310. http://doi.org/10.1080/09588221.2013.839569

Chapelle, C. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge: Cambridge University Press.

Chapelle, C. A., & Sauro, S. (2017). *The Handbook of Technology and Second Language Teaching and Learning*. Hoboken: Wiley-Blackwell.

Chen, J. C. (2016). The crossroads of English language learners, task-based

instruction, and 3D multi-user virtual learning in Second Life. *Computers & Education*, 102, pp. 152–171. http://doi.org/10.1016/j.compedu.2016.08.004

Cheon, H. (2003). The viability of computer mediated communication in the Korean secondary EFL classroom. *Asian EFL Journal*, *5*(1), pp. 1-61.

Chik, A. (2014). Digital gaming and language learning: Autonomy and community. *Language Learning & Technology*, *18*(2), pp. 85–100

Choi, H. (2006). *A study on flow element in MMORPG*. Master, Kookmin, Seoul.

Chun, D. M. (2006). CALL technologies for L2 reading. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching*. San Marcos, TX: CALICO, pp. 81–98.

Cohen, L., Manion, L. & Morrison K. (2000). *Research Methods in Education*. (5th ed.) London: Routledge.

Cornillie, F., Thorne, S. L., & Desmet, P. (2012). ReCALL Special issue: Digital games for language learning: challenges and opportunities. *ReCALL*, *24*(3), pp. 243-256.

Cornillie, F., Clarebout, G., & Desmet, P. (2012). Between learning and playing? Exploring learners' perceptions of corrective feedback in an immersive game for English pragmatics. *ReCALL*, *24*(3), pp. 257–278. http://doi.org/10.1017/S0958344012000146

Coughlan, P., & Duff, P.A. (1994). Same task, different activities: analysis of a SLA task froman activity theory perspective. In J. Lantol and G. Appel (Eds.), *Vygotskian approaches to second language research*. Norwood, NJ: Ablex, pp. 173–94.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Crookes, G. (1986). *Task classification: A cross-disciplinary review* (No. 4). Center for Second Language Classroom Research, Social Science Research Institute, University of Hawaii at Manoa.

Darhower, M. (2002). Interactional features of synchronous computer-mediated communication in the intermediate L2 class: A sociocultural case study. *CALICO Journal*, *19*(2), pp. 249–278.

Darhower, M. A. (2008). The Role of Linguistic Affordances in Telecollaborative Chat. *CALICO Journal*, *26*(261), 48–69. https://doi.org/10.1558/cj.v26i1.48-69

Darhower, M. (2013). Interactional Features of Synchronous Computer-Mediated Communication in the Intermediate L2 Class: A Sociocultural Case Study. *CALICO Journal*, *19*(2), 249–277. https://doi.org/10.1558/cj.v19i2.249-277

De Marco, A., & Leone, P. (2013). Discourse Markers in Italian as L2 in face to face vs.

computer mediated settings. In Bradley, L. & S. Thouësny (Eds.), *20 years of EUROCALL: Learning from the past, looking to the future*. Dublin: Researchpublishing.net, pp. 71-77.

deHaan, J. W. (2005a) Learning Language through Video Games: A Theoretical Framework, An Evaluation of Game Genres and Questions for Future Research. In: S. P. Schaffer, & M. L. Price, (Eds.), *Interactive convergence: critical issues in multimedia*. Oxford: InterDisciplinary Press, pp. 229–239. http://www.inter-disciplinary.net/publishing/id-press/ebooks/interactive-convergence-critical-issues-in-multimedia/

deHaan, J. W. (2005b). Acquisition of Japanese as a foreign language through a baseball video game. *Foreign Language Annals, 38*(2), pp. 278–282.

deHaan, J. W. & Diamond, J. (2007). The experience of telepresence with a foreign language video game and video. *Proceedings of the 2007 ACM SIGGRAPH symposium on Video games*, pp. 39–46.

deHaan, J. W. (2008). *Video games and second language acquisition: the effect of interactivity with a rhythm video game on second language vocabulary recall, cognitive load, and telepresence*. Ph.D. thesis, New York University.

deHaan, J. W., & Kono, F. (2010). The effect of interactivity with WarioWare Minigames on second language vocabulary learning. *Journal of Digital Games Research*, *4*(2), pp. 47-59.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pp. 9-15.

Develotte, C. (2009). From face to face to distance learning: the online learner's emerging identity. In R. Goodfellow & M. N. Lamy (Eds.), *Learning Cultures in Online Education*. London: Continuum, pp. 71-92.

Dieterle, E., & Clarke, J. (2008). Multi-user virtual environments for teaching and learning. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking*. Hershey, PA: Idea Group, pp. 1033-1041.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction*. Cambridge: Cambridge University Press, pp. 206-257.

Doughty, C., & Williams, J. (Eds.). (1998). *Focus on form in classroom second language acquisition*. New York: Cambridge University Press.

Duquette, J. P. (2013). Cypris Chat: Emergent Design for Language Education in Second Life. *In 2013 International Conference on Informatics and Creative Multimedia*. IEEE, pp. 70-75. https://doi.org/10.1109/icicm.2013.21

DuQuette, J., & Hann, F. (2010). Using tasks in virtual worlds. *Temple University Japan: Working Papers in Applied Linguistics*, *23*, pp. 19-26.

Economic and Social Research Council (ESCR). (2015). *ESRC Framework for research ethics Updated January 2015*. Available at: https://esrc.ukri.org/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/ (Accessed: 25 January 2016).

Ejezenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency*. Ann Arbor: University of Michigan Press, pp. 287-314.

Ellis, D. (2011). The role of task complexity in the linguistic complexity of native speaker output. *Qualifying paper, PhD in Second Language Acquisition Program.* University of Maryland.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University.

Ellis, R. (2000). Task-based research and language pedagogy. *Language teaching research*, *4*(3), pp. 193-220.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. (2005). Planning and task-based performance: Theory and research. In R. Ellis (Ed.), *Planning and task performance in a second language*. Philadelphia: John Benjamins, pp. 3-34.

Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies*. Oxford: Oxford University Press, pp. 339-360.

Ellis, R. (2009). The Differential Effects of Three Types of Task Planning on the Fluency, Complexity, and Accuracy in L2 Oral Production. *Applied Linguistics*, *30*(4), pp. 474-509.

Ellis, R. & Barkhuizen, G. (2005). *Analyzing Learner Language*. Oxford: Oxford University Press.

Filsecker, M., & Bündgens-Kosten, J. (2012). Behaviorism, Constructivism, and Communities of Practice: How pedagogic theories help us understand game-based language learning. In *Digital games in language learning and teaching*. London: Palgrave Macmillan, pp. 50-69.

Fitze, M. (2006). Discourse and participation in ESL face-to-face and written electronic conferences. *Language Learning and Technology*, *10*(1), pp. 67–86.

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based

learning. *Studies in Second Language Acquisition*, *18*(3), pp. 299–324.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), pp. 354–375.

Franciosi, S. J. (2015). Acceptability of RPG simulators for foreign language training in Japanese higher education. *Simulation & Gaming*, *47*(1), pp. 31-50.

Furstenberg, G., and Malone, S. (1993) *A la rencontre de Philippe*, New Haven: Yale University Press.

Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral production. *Language Teaching Research*, *20*(3), pp. 321–340.

Gánem-Gutiérrez, G. A. (2014). The Third Dimension: a Sociocultural Theory Approach to the Design and Evaluation of 3D Virtual Worlds Tasks.' In M. Gonzalez-Lloret, & L. Ortega. (Eds.) *Technology and Tasks: Exploring Technology-Mediated TBLT. Task-Based Language Learning*. Amsterdam: John Benjamins Publishing Company, pp. 213-238.

Garrett, N. (2009). Computer-assisted language learning trends and issues revisited: Integrating innovation. *The Modern Language Journal*, *93*(s1), pp. 719-740.

Gass, S. and Mackey, A. (2007). Input, interaction, and output in second language acquisition. In B. VanPatten and J. Williams (Eds*.), Theories in second language acquisition: An introduction*. Mahwah, NJ: Lawrence Erlbaum, pp. 175–200.

Gaudart, H. (1999). Games as Teaching Tools for Teaching English to Speakers of Other Languages. *Simulation & Gaming*, *30*(3), pp. 283–291.

Gee, J. P. (2004). *Situated language and learning: A critique of traditional schooling*. London: Routledge.

Gee, J. P. (2005). Semiotic social spaces and affinity spaces: From the age of mythology to today's schools. In D. Barton & K. Tusting (Eds.), *Beyond communities of practice: Language, power, and social context* (pp. 214–232). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511610554.012

Gee, J. (2007). *Good video games and good learning*. New York: Peter Lang.

Gee, J. P., & Hayes, E. (2012). Nurturing affinity spaces and game-based learning. *Games, learning, and society: Learning and meaning in the digital age*, *123*, pp. 1-40.

George, A., & Sanders, M. (2017). Evaluating the potential of teacher-designed technology-based tasks for meaningful learning: Identifying needs for professional development. *Education and Information Technologies*, *22*(6), 2871–2895. http://doi.org/10.1007/s10639-017-9609-y

Geng, X., & Ferguson, G. (2013). Strategic planning in task-based language teaching: The effects of participatory structure and task type. *System*, *41*(4), pp. 982–993.

Gibson, C. (n.d.). *Digital Dialects Japanese Hiragana*. Available at: http://www.digitaldialects.com/Japanese/Hiragana.htm (Accessed: 10 May 2015).

Gilabert, R. (2006). The simultaneous manipulation of task complexity along planning time and here-and-now: Effects on L2 oral production; Investigating tasks in formal language learning; second language acquisition. In M.P. García Mayo (Ed.), *Investigating tasks in formal language learning*. Clevedon, England: Multilingual Matters, pp. 44–68.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics*, 45(2), pp. 215-240.

Gilabert, R., Barón, J., & Llanes, A. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *International Review of Applied Linguistics*, *47*(3-4), pp. 365-395.

González-Lloret, M. (2014). The need for needs analysis in technology-mediated TBLT. In M. Gonzalez-Lloret & L. Ortega (Eds.). *Technology-mediated TBLT: Researching technology and tasks*. Amsterdam: John Benjamins, pp. 23-50.

González-Lloret, M. (2015a). *A Practical Guide to Integrating Technology into Task-Based Language Teaching*. Washington DC: Georgetown University Press.

González-Lloret, M. (2015b). Conversation analysis in computer-assisted language learning. *CALICO Journal*, 32(3), pp. 569–594.

González-Lloret, M., & Ortega, L. (2014). Towards technology-mediated TBLT: An introduction. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks*. Amsterdam: John Benjamins, pp. 1–22.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), pp. 123-145.

Gray, E. F. (1992). Interactive Language Learning: A la rencontre de Philippe. *The French Review*, *65*(3), pp. 499–507.

Guara-Tavares, M. G. 2008. *Pre-task Planning, Working Memory Capacity and L2 Speech Performance.* Unpublished Doctoral Thesis, Universidade Federal de Santa Catarina, Brazil.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Paris: Presses Universitaires de France

Gygax, G., & Arneson, D. (1974). *Dungeons and dragons Vol. 19*. Lake Geneva, WI: Tactical Studies Rules.

Hammersley, M. (2010). *Creeping Ethical Regulation and the Strangling of Research*. Sociological Research Online, 15(4), pp. 1–3. https://doi.org/10.5153/sro.2255

Hampel, R. (2006). Rethinking task design for the digital age: A framework for language teaching and learning in a synchronous online environment. *ReCALL*, *18*(1), pp. 105-121. http://doi.org/10.1017/S0958344006000711

Hampel, R. (2010). Task design for a virtual learning environment in a distance language course. In M. Thomas & H. Reinders (Eds.), *Task-Based Language Learning and Teaching with Technology*. London: Continuum, pp. 131-153.

Hansen, S., Berente, N., Pike, J., & Bateman, P. J. (2008). Productivity and Play in Organizations: Executive perspectives on the real-world organizational value of immersive virtual environments. *Artifact*, 2(2), pp. 69-81.

Harris, K. (2005). Same activity, different focus. *Focus on Basics*, 8(1), pp. 7-10.

Hastings, C. R. (2014). Social Learning Spaces: An Investigation into the Effects of Using Collaborative Strategic Board Games as Modes of Oral Practice in Combination with Explicitly Teaching Interactional Competence on the Oral Proficiency and Discourse of Japanese Learners. *Unpublished MA Thesis for Oxford Brookes*.

Hawkes, M. L. (2011). Using task repetition to direct learner attention and focus on form. *ELT Journal*, *66*(3), pp. 327–336.

Hayes, C. A., & Holmevik, J. R. (2001). *High wired: On the design, use, and theory of educational MOOs*. Ann Arbor, MI: University of Michigan Press.

Henderson, M., Huang, H., Grant, S., & Henderson, L. (2012). The impact of Chinese language lessons in a virtual world on university students' self-efficacy beliefs. *Australasian Journal of Educational Technology*, *28*(3), pp. 400–419. doi:10.1007/BF01172995

Higgins, J. (1983) Computer Assisted Language Learning, *Language Teaching*, 16, pp. 102-114.

Hourdequin, P. York, J. & deHaan, J. (2017) Learning English and Other 21st Century Skills Through Games: Lessons for Japanese Higher Education from Learning Spaces in New York City. *Tokoha Gakuen University Research Review Faculty of Foreign Studies* 33, pp. 41-59.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), pp. 461–473.

Housen, A., Kuiken, F., & Vedder, I. (2012). Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA. Language Learning & Language Teaching. Volume 32. *Language Learning & Language Teaching (MS)*, *8*, pp. 269-274.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational researcher*, *17*(8), pp. 10-16.

Hung, Y. W., & Higgins, S. (2016). Learners' use of communication strategies in text-based and video-based synchronous computer-mediated communication environments: opportunities for language learning. *Computer Assisted Language Learning*, *29*(5), pp. 901–924. http://doi.org/10.1080/09588221.2015.1074589

Hung, H. T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C. C. (2018). A scoping review of research on digital game-based language learning. *Computers & Education*, 126, pp. 89-104.

Hunt, K. (1965). Grammatical structures written at three grade levels. *NCTE Research Report*, 3. Champaign, Ilinois: NCTE.

Iwasaki, J., & Oliver, R. (2003). Chat-line interaction and negative feedback. *Australian Review of Applied Linguistics*, *17*(1), pp. 60–73.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can We Predict Task Difficulty in an Oral Proficiency Test? Exploring the Potential of an Information-Processing Approach to Task Design. *Language Learning*, *51*(3), pp. 401–436. https://doi.org/10.1111/0023-8333.00160

Jarvis, S., & Daller, M. (2013). *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: John Benjamins Publishing.

Jauregi, K., Canto, S., de Graaff, R., Koenraad, T., & Moonen, M. (2011). Verbal interaction in Second Life: towards a pedagogic framework for task design. *Computer Assisted Language Learning*, *24*(1), pp. 77–101. http://doi.org/10.1080/09588221.2010.538699

Jauregi, K., Kuure, L., Bastian, P., Reinhardt, D., & Koivisto, T. (2015). Cross-cultural discussions in a 3D virtual environment and their affordances for learners' motivation and foreign language discussion skills. *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, pp. 274-280.

Jackson, D. O., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A Synthesis and Meta-Analysis of Research on Second Language Task Complexity. *Language Learning*, *63*(2), pp. 330-367.

Jee, M. J. (2014). From First Life to Second Life: Evaluating task-based language learning in a new environment / De la vie réelle à la vie virtuelle: évaluation de l'apprentissage des langues basé sur les tâches dans un nouvel environnement. *Canadian Journal of Learning and Technology / La Revue Canadienne de L'apprentissage et de La Technologie; Vol 40, No 1 (2014)*, *40*(1), pp. 1-15.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, 24(4), pp. 602-611.

Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K. (2011). *The 2011 Horizon Report*. Austin, TX: The New Media Consortium.

Johnson, K., & Johnson, H. (Eds). (1999). *An encyclopedic dictionary of applied linguistics: A handbook for language teaching*. Malden, MA: Blackwell Publishers.

Johnson, R. B., & Onquegbuzie, A.J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational researcher*, *33*(7), pp. 14-26.

Kalyuga, S., & Plass, J. L. (2009). Evaluating and managing cognitive load in games. In *Handbook of research on effective electronic gaming in education*. Hershey, PA: Information Science Reference, pp. 719-737.

Kapp, K. (2012). *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*. San Francisco, CA: Pfeiffer

Kim, H. (2017). Effect of modality and task type on interlanguage variation. ReCALL, 29(2), 219-236. doi:10.1017/S0958344017000015

Kim, Y. (2009). The effects of task complexity on learner-learner interaction. *System*, *37*(2), pp. 254-268.

Klapper, J. (2003). Taking communication to task. A critical review of recent trends in language teaching. *Language Learning Journal*, *27*(1), pp. 33-42.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *1*(1), pp. 60-69. http://doi.org/10.7820/vli.v01.1.koizumi

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), pp.145-164.

Kormos, J., & Trebits, A. (2012). The Role of Task Complexity, Modality, and Aptitude in Narrative Task Performance. *Language Learning*, *62*(2), pp. 439-472. https://doi.org/10.1111/j.1467-9922.2012.00695.x

Kramsch, C. and Thorne, S. L. (2002) Foreign language learning as global communicative practice. In: Block, D. and Cameron, D. (eds.), *Globalization and language teaching.* London: Routledge, 83–100.

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*. Philadelphia/Amsterdam: John Benjamins, pp. 91–104.

Kuiken, F., Vedder, I., (2006). Cognitive task complexity and linguistic performance in French L2 writing. In M. P. Garcia Mayo (Ed.), *Investigating tasks in formal language learning*. New York: Multilingual Matters Ltd., pp. 117-135.

Kuhn T.S. (1970) *The Structure of Scientific Revolutions*. Chicago University Press: Chicago.

Kuhn, J., & Stevens, V. (2017). Participatory culture as professional development: Preparing teachers to use Minecraft in the classroom. *TESOL Journal*, *8*(4), pp. 753-767. http://doi.org/10.1002/tesj.359

Lambert, C.P., and Engler, S. (2007). Information Distribution and Goal Orientation in Second Language Task Design. In: García Mayo, M.d.P. (Ed.), *Investigating tasks in formal language learning*. New York: Multilingual Matters Ltd., pp. 27-43.

Lan, Y. J. (2014). Does Second Life improve Mandarin learning by overseas Chinese students. *Language Learning & Technology*, *18*(2), pp. 36-56.

Landwehr, P., Diesner, J., & Carley, K. M. (2009). Words of Warcraft: a relational text analysis of quests in an MMORPG. *Proceedings of Digital Games Research Association Conference (DiGRA)*. London: UK.

Larsen-Freeman, D. (2012). On the roles of repetition in language teaching and learning. *Applied Linguistics Review*, *3*(2), pp. 195-210.

Lee, L. (2002). Enhancing learners' communication skills through synchronous electronic interaction and task-based instruction. *Foreign Language Annuals*, *35*(1), pp. 16-23.

Lee, L. (2016). Autonomous learning through task-based instruction in fully online language courses. *Language Learning & Technology*, *20*(2), pp. 81-97.

Lee, Y.J., & Gerber, H. (2013). It's a WOW world: Second language acquisition and massively multiplayer online gaming. *Multimedia-Assisted Language Learning*, *16*(2), pp. 53-70.

Lee, J.Y., & Pass, C. (2014). Massively multiplayer online gaming and English language learning. In H.R. Gerber & S.S. Abrams (Eds.), *Bridging literacies with videogames*. Rotterdam: Sense Publishers, pp. 91-101.

Lennon, P. 1990. Investigating fluency in EFL: a quantitative approach. *Language Learning, 40*(3) pp. 387-417.

Levak, N., & Son, J. B. (2017). Facilitating second language learners' listening comprehension with Second Life and Skype. *ReCALL*, *29*(2), pp. 1-19. http://doi.org/10.1017/S0958344016000215.

Levkina, M., & Gilabert, R. (2012). The effects of cognitive task complexity on L2 oral production. In Housen, A., Kuiken, F., & Vedder, I. (Eds.), *Dimensions of L2 performance and proficiency: investigating complexity, accuracy, and fluency in SLA* (pp. 171–198). Amsterdam, TheNetherlands: John Benjamins.

Lim, C. P., Nonis, D., & Hedberg, J. (2006). Gaming in a 3D multiuser virtual

environment: Engaging students in science lessons. *British Journal of Educational Technology*, *37*(2), pp. 211-231.

Lin, H. (2015). Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis. *ReCALL*, *27*(3), pp. 261–287.

Linden, J. (2018). Build Tools. Available at: https://community.secondlife.com/knowledgebase/english/build-tools-r12/ (Accessed: 26 August 2018).

Lintunen, P., & Mäkilä, M. (2014). Measuring Syntactic Complexity in Spoken and Written Learner Language: Comparing the Incomparable? *Research in Language*, *12*(4), p. 377-399. https://doi.org/10.1515/rela-2015-0005

Liou, H.C. (2012). The roles of Second Life in a college computer assisted language learning (CALL) course in Taiwan, ROC. *Computer Assisted Language Learning*, *25*(4), pp. 365-382.

Lombardi, I. (2015). Fukudai Hero: A Video Game-like English Class in a Japanese National University. *EL.LE: Educazione Linguistica. Language Education*, *4*(3), pp. 483-499. http://doi.org/10.14277/2280-6792/ELLE-4-3-15-7

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie and T.K. Bhatia (Eds): *Handbook of Second Language Acquisition*. New York: Academic Press, pp. 413–68.

Long, M. H. (2005). Methodological issues in learner needs analysis. In M. H. Long (Ed.), *Second language needs analysis*. Cambridge, UK: Cambridge University Press, pp. 1-76.

Long, M. (2012). Focus on form in task-based language teaching. Cited from Ryu, D. H: Focus-on-form revisited: Using 17 again. STEM Journal, *13*(1), pp. 19-39.

Long, M. (2014). *Second language acquisition and task-based language teaching*. Malden, MA: John Wiley & Sons.

Long, M. (2016). In Defense of Tasks and TBLT: Nonissues and Real Issues. Annual Review of *Applied Linguistics*, 36, 5-33. doi:10.1017/S0267190515000057

Lord, G. (2008) Podcasting communities and second language pronunciation. *Foreign Language Annals*, *41*(2), pp. 374-389.

Lund, A., & Lund, M. (2015). One-way repeated measures ANOVA in SPSS statistics. Available at: *https://statistics.laerd.com/premium/rma/repeated-measures-anova-in-spss.php.* (Accessed: January 31 2018).

Lyster, R. and Ranta, L. (1997). Corrective feedback and learner uptake: negation of form in communicative classrooms. *Studies in Second Language Acquisition*, *19*(1), pp. 37-

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied linguistics*, *27*(3), pp. 405-430.

Mackey, A., Abbuhl, R., & Gass, S. M. (2012). Interactionist approach. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition*. New York, NY: Routledge, pp. 7-23.

Malone, T. W. (1981). What makes computer games fun? *BYTE*, 5, 258-277.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan and A. Wray (Eds.), *Evolving models of language*. Clevedon: Multilingual Matters, pp. 58–71.

Malvern, D.D., Richards, B.J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Hampshire: Palgrave Macmillan.

Marion, R. (2004). *The whole art of deduction: Research skills for new scientists.* University of Texas Medical Branch. Retrieved from: http://www.sahs.utmb.edu/pellinore/intro_to_research/wad/sel_test.htm

Marklund, B., & Taylor, A. S. (2015). Teachers' many roles in game-based learning projects. In *European Conference on Games Based Learning 2015* (pp. 359-367). Academic Conferences and Publishing International Limited.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), pp. 381–92. http://doi.org/10.3758/BRM.42.2.381

McDonough, K. (2005). Identifying the impact of negative feedback and learners' response on ESL question development. *Studies in Second Language Acquisition*, *27*(1), pp. 79-103.

McDonough, K. (2007). Interactional feedback and the emergence of simple past activity verbs in L2 English. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies*. Oxford: Oxford University Press, pp. 323-338.

Melchor-Couto, S. (2017). Foreign language anxiety levels in Second Life oral interaction. *ReCALL, 29*(1), pp. 99-119.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, *45*(3), pp. 241-259.

Miller, M., & Hegelheimer, V. (2006). The SIMs meet ESL Incorporating authentic computer

simulation games into the language classroom. *Interactive Technology and Smart Education*, *3*(4), pp. 311-328. http://doi.org/10.1108/17415650680000070

Milton, J., Jonsen, S., Hirst, S., & Lindenburn, S. (2012). Foreign language vocabulary development through activities in an online 3D environment. *The Language Learning Journal*, *40*(1), pp. 99-112.

Ministry of Education, Culture, Sports, Science, and Technology. (2014). *On Integrated Reforms in High School and University Education and University Entrance Examination Aimed at Realizing a High School and University Articulation System Appropriate for a New Era*. Available at: http://www.mext.go.jp/en/news/topics/detail/1372628.htm (Accessed: 17 March 2015).

Molin, G. (2017). The role of the teacher in game-based learning: A review and outlook. *Serious Games and Edutainment Applications: Volume II*, 649–674. https://doi.org/10.1007/978-3-319-51645-5_28

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics 38* (8): pp. 114–117.

Nardi, B. (2010). *My life as a night elf priest: An anthropological account of World of Warcraft*. Ann Arbor: University of Michigan Press.

Newgarden, K., Zheng, D. P. and Liu, M. (2015) An eco-dialogical study of second language learners' World of Warcraft (WoW) gameplay. *Language Sciences,* 48, pp.22-41.

Newgarden, K., & Zheng, D. (2016). Recurrent languaging activities in World of Warcraft: Skilled linguistic action meets the Common European Framework of Reference. *ReCALL*, *28*(3), pp. 274-304. http://doi.org/10.1017/S0958344016000112

Nicholson S. (2015) A RECIPE for Meaningful Gamification. In: Reiners T., Wood L. (eds) *Gamification in Education and Business*. Springer, Cham

Nik, N. (2010). *Examining the language learning potential of a task-based approach to synchronous computer-mediated communication* (Unpublished doctoral dissertation). Victoria University of Wellington, New Zealand.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta–analysis. *Language Learning, 50*(3), pp. 417-528.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), pp. 555-578. http://doi.org/10.1093/applin/amp044

Nunan, D. (2004). *Task-based language teaching.* Cambridge: Cambridge University Press.

OECD. (2015). *Students, computers and learning: Making the connection*. PISA: OECD

Publishing.

Oskoz, A., & Elola, I. (2014). Promoting foreign language collaborative writing through the use of Web 2.0 tools. In M. González-Lloret & L. Ortega (Eds.), *Technology and Tasks: Exploring Technology- mediated TBLT*. Philadelphia, PA: John Benjamins, pp. 115-148.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), pp. 429-434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79, pp. 419-430.

Paharia, R. (2010). Who Coined the Term ”Gamification”? Available at: http://www.quora.com/Who-coined-the-term-gamification (Accessed: April 27 2013).

Park, M. (2018). Innovative assessment of aviation English in a virtual world: Windows into cognitive and metacognitive strategies. *ReCALL*, *30*(2), 196–213. https://doi.org/10.1017/S0958344017000362

Peachey, A., Gillen, J., Livingstone, D., & Smith-Robbins, S. (Eds.). (2010). *Researching learning in virtual worlds*. Springer Science & Business Media.

Pellettieri, J. (2000). Negotiation in cyberspace: The role of chatting in the development of grammatical competence. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice*. Cambridge: Cambridge University Press, pp. 59-86.

Peterson, M. (2005). Learning interaction in an avatar-based virtual environment: a preliminary study. *PacCALL Journal*, *1*(1), pp. 29- 40.

Peterson, M. (2006). Learner interaction management in an avatar and chat-based virtual world. *Computer Assisted Language Learning*, *19*(1), pp. 79-103. doi:10.1080/09588220600804087

Peterson, M. (2009). Learner interaction in synchronous CMC: a sociocultural perspective. *Computer Assisted Language Learning*, *22*(4), pp. 303–321. http://doi.org/10.1080/09588220903184690

Peterson, M. (2010a). Task-based language teaching in network-based CALL: An analysis of research on learner interaction in synchronous CMC. In M. Thomas & H. Reinders (Eds.), *Task-Based Language Teaching and Technology*. London: Continuum, pp. 41-62.

Peterson, M. (2010b). Learner participation patterns and strategy use in Second Life: an exploratory case study. *ReCALL*, 22(03), pp.273-292.

Peterson, M. (2012). Learner interaction in a massively multiplayer online role playing game (MMORPG): A sociocultural discourse analysis. *ReCALL*, *24*(03), pp. 361-380. doi:10.1017/S0958344012000195

Peterson, M. (2016a). *Computer games and language learning*. New York, NY: Springer.

Peterson, M. (2016). The use of massively multiplayer online role-playing games in CALL: an analysis of research. *Computer Assisted Language Learning*, *29*(7), pp. 1181-1194. http://doi.org/10.1080/09588221.2016.1197949

Philp, J. (2003). Constraints on "noticing the gap": Non-native speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition*, *25*(1), pp. 99–126.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction. In G. Crookes & S. Gass (Eds.), *Tasks and Language Learning: Integrating Theory and Practice. Vol 1*. Clevedon, England: Multilingual Matters, pp. 9-34.

Pica, T., Kang, H. S., & Sauro, S. (2006). Information gap tasks: Their multiple roles and contributions to interaction research methodology. *Studies in Second Language Acquisition*, *28*(2), pp. 301-338. http://doi.org/10.1017/S027226310606013X

Rahimpour, M. (1999). Task complexity and variation in interlanguage. In N. O. Jungheim & P. Robinson (Eds.), *Pragmatics and Pedagogy: Proceedings of the 3rd Pacific Second Language Research Forum*. Tokyo, Japan: The Pacific Second Language Research Forum, pp. 115-134.

Rama, P. S., Black, R. W., van Es, E., & Warschauer, M. (2012). Affordances for second language learning in World of Warcraft. *ReCALL*, *24*(3), pp. 322-338. http://doi.org/10.1017/S0958344012000171

Rankin J. (1995). The effects of task design on accuracy and self-monitoring. American Association of Applied Linguistics. cf: Lambert, C.P., and Engler, S. (2006). Information Distribution and Goal Orientation in Second Language Task Design. In: García Mayo, M.P. (2006). *Investigating tasks in formal language learning*. New York: Multilingual Matters.

Rankin, Y., Gold, R., & Gooch, B. (2006). Evaluating Interactive Gaming as a Language Learning Tool. *ACM SIGGRAPH 2006 Educators Program*, pp. 1-6. https://doi.org/10.1145/1179295.1179340

Rankin, Y., McNeal, M., Shute, M. W., & Gooch, B. (2008). User centered game design: evaluating massive multiplayer online role playing games for second language acquisition. In *Proceedings of the 2008 ACM SIGGRAPH symposium on Video games*. ACM, pp. 43-49.

Reinders, H., & Wattana, S. (2011). Learn English or die: The effects of digital games on interaction and willingness to communicate in a foreign language. *Digital Culture &*

Education, 3(1), pp. 4-28.

Reinders, H., & Wattana, S. (2014). Can I say something? The effects of digital game play on willingness to communicate. *Language Learning & Technology*, 18(2), pp. 101-123.

Reinders, H., & Wattana, S. (2015). Affect and willingness to communicate in digital game-based learning. *ReCALL*, *27*(1), pp. 38-57. http://doi.org/10.1017/S0958344014000226

Reinhardt, J. (2008). Negotiating meaningfulness: An enhanced perspective on interaction in computer-mediated foreign language learning environments. In S. Magnan (Ed.), *Mediating Discourse Online*. Amsterdam: Benjamins, pp. 219-244.

Reinhardt, J., & Sykes, J. M. (2012). Conceptualizing digital game-mediated L2 learning and pedagogy: Game-enhanced and game-based research and practice. In *Digital games in language learning and teaching*. London: Palgrave Macmillan, pp. 32-49.

Reinhardt, J., & Sykes, J. M. (2014). Digital game and play activity in L2 teaching and learning. *Language Learning and Technology*, *18*(2), pp. 2-8.

Reinhardt, J., & Thorne, S. (2016). Metaphors for digital games and language learning. In F. Farr & L. Murray (Eds.), *The Routledge Handbook of Language Learning and Technology*. Oxford: Routledge, pp. 415-430.

Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom-based study. *The Modern Language Journal*, 95, pp. 162-181.

Révész, A., Ekiert, M., & Torgersen, E. N. (2014). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics 37*(6), pp. 828-848.

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition*, *38*(4), pp. 703-737.

Robinson, P. (1995): Task complexity and second language narrative discourse. *Language Learning*, *45*(1), pp. 99-140.

Robinson, P. (2001a). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics*, *22*(1), pp. 27-57.

Robinson, P. (2001b). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), Cognition and second language instruction. Cambridge, UK: Cambridge University Press, pp. 287-318.

Robinson, P. (2007). Criteria for grading and sequencing pedagogic tasks. In M. P. Garcia Mayo (Ed.), *Investigating Tasks in Formal Language Learning*. Clevedon:

Multilingual Matters, pp. 7-27.

Robinson, P. (2009). Syllabus Design. In M. H. Long and C. J. Doughty (Eds.), *The Handbook of Language Teaching*. New Jersy: Blackwell, pp. 294-310.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.). *Second language task complexity: Researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins, pp. 3-37.

Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(3), pp. 161-176. http://doi.org/10.1515/iral.2007.007

Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, 30(4), pp. 533-554.

Roed, J. (2003). Language Learner Behaviour in a Virtual Environment. *Computer Assisted Language Learning: An International Journal*, 16(2-3), pp. 155-72.

Ryle, A. (1999). Object relations theory and activity theory: a proposed link by way of the procedural sequence model. In Y. Engeström, R. Miettinen & R. Punamäki (Eds.), *Perspectives on activity theory*. Cambridge: Cambridge University Press, pp. 407-418.

Sajavaara, K., & Lehtonen, J. (1978). Spoken language and the concept of fluency. *Language Centre News*, 1, pp. 23-57.

Sample, E., & Michel, M. (2014). An Exploratory Study into Trade-off Effects of Complexity, Accuracy, and Fluency on Young Learners' Oral Task Repetition. *TESL Canada Journal*, *31*(8), pp. 23-46.

Samuda, V., & Bygate, M. (2008). *Tasks in Second Language Learning*. Palgrave Macmillan.

Sasayama, S., & Izumi, S. (2012). Effects of task complexity and pre-task planning on EFL learners' oral production. In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation*. Amsterdam, Netherlands: John Benjamins, pp. 23–42.

Sasayama, S. (2013). *Is a 'complex' task really complex? Measuring task complexity independently from linguistic production*. Paper presented at the 5th Biennial International Conference on Task-Based Language Teaching, Banff, Alberta, Canada.

Sasayama, S. (2015). *Validating the assumed relationship between task design, cognitive complexity, and second language task performance*. Georgetown University. Unpublished PhD Thesis.

Satar, H., & Özdener, N. (2008). The effects of synchronous CMC on speaking proficiency and anxiety: Text versus voice chat. *The Modern Language Journal*, *92*(4), pp. 595-613.

Sato, R. (2010). Reconsidering the effectiveness and suitability of PPP and TBLT in the Japanese EFL classroom. *JALT Journal* 32(2): pp. 189-200.

Sauro, S. (2011). SCMC for SLA: A research synthesis. *CALICO Journal*, 28(2), pp. 369-391.

Sauro, S. (2014). Lessons from the fandom: Task models for technology-enhanced language learning. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks*. Philadelphia, PA: John Benjamins, pp. 239-262.

Sauro, S., & Sundmark, B. (2016). Report from Middle-Earth: fan fiction tasks in the EFL classroom. *ELT Journal*, *70*(4), pp. 414-423. http://doi.org/10.1093/elt/ccv075

Schmidt, R. (1990). The role of consciousness is second language learning. *Applied Linguistics*, 11, pp. 129-158.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *Consciousness in second language learning*, *11*, pp. 237-326.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human - Computer Studies, 74*, 14-31. doi:10.1016/j.ijhcs.2014.09.006

Sega AM2. (1999). *Shenmue*. Dreamcast: Sega.

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research*, *10*(4), pp. 361–392.

Sheldon, L. (2011). *The multiplayer classroom: Designing coursework as a game*. Boston, MA: Cengage Learning.

Shetzer, H., & Warschauer, M. (2000). An electronic literacy approach to network-based language teaching. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice*. Cambridge: Cambridge University Press, pp. 171-185.

Skehan, P. (1991). Individual differences in second language learning. *Studies in second language acquisition*, *13*(2), pp. 275-298.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan

& M. Swain (Eds.), *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Harlow: Pearson Education, pp. 167-185.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), pp. 1-14.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), pp. 510-532.

Skehan, P. (2016). Tasks Versus Conditions: Two Perspectives on Task Research and Their Implications for Pedagogy. *Annual Review of Applied Linguistics*, 36, pp. 34-49.

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task based learning. *Language Teaching Research, 1*(3), pp. 185-211.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), pp. 93-120.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.). *Cognition and second language instruction*. Cambridge: Cambridge University Press, pp. 183-205.

Smart, J., Cascio, J. & Paffendof, J. (2007). Metaverse roadmap: pathways to the 3D web. Available at: http://www.metaverseroadmap.org/overview (Accessed: August 18, 2018).

Smith, B. (2003). Computer-mediated negotiated interaction: An expanded model. *The Modern Language Journal*, *87*(1), pp. 38-57.

Smith, B. (2005). English language learning and technology: Lectures on applied linguistics in the age of information and communication technology. *Modern Language Journal*, *89*(4), pp. 647-648.

Steinkuehler, C.A., 2006. Massively multiplayer online video gaming as participation in a discourse. *Mind, Culture, and Activity*, *13*(1), pp. 38-52.

Stockwell, G. (2007). A review of technology choice for teaching language skills and areas in the CALL literature. *ReCALL*, *19*(2), pp. 105-120.

Suh, S., Kim, S. W., & Kim, N. J. (2010). Effectiveness of MMORPG-based instruction in elementary English education in Korea. *Journal of Computer Assisted Learning*, *26*(5), pp. 370-378.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of henry G. Widdowson*. Oxford: Oxford University Press, pp. 125-144.

Swain, M. (2006). Languaging, agency and collaboration in advanced language proficiency. In H. Byrnes (Ed.), *Advanced Language Learning: The Contribution of*

Halliday and Vygotsky. London: Continuum, pp. 95-108.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, *82*(3), pp. 320-337.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312.

Swier, R. (2014). Tasks for easily modifiable virtual environments. *JALT CALL Journal*, *10*(3), pp. 203-219.

Sykes, J. M., Oskoz, A., & Thorne, S. L. (2008). Web 2.0, synthetic immersive environments, and mobile resources for language education. *Calico Journal*, *25*(3), pp. 528-546.

Sykes, J. M. (2014). TBLT and synthetic immersive environments: What can in-game task restarts tell us about design and implementation? In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks*. Amsterdam, The Netherlands: John Benjamins, pp. 149-182.

Sykes, J. M. (2018). Digital games and language teaching and learning. *Foreign Language Annals*, *51*(1), pp. 219-224.

Sykes, J. M., & Reinhardt, J. (2013). *Language at play: Digital games in second and foreign language teaching and learning*. Boston, MA: Pearson.

Takahashi, D. (2010). *Gamification gets its own conference*. Available at: https://venturebeat.com/2010/09/30/gamification-gets-its-own-conference/ (Accessed: August 23, 2018).

Thomas, M. (2009). *Handbook of research on Web 2.0 and second language learning*. Hershey, PA: Information Science Reference.

Thomas, M., & Reinders, H. (2010). *Task-based language learning and teaching with technology*. Continuum International Pub Group.

Thorne, S. L., & Black, R. W. (2007). Language and literacy development in computer-mediated contexts and communities. *Annual Review of Applied Linguistics*, *27*, pp. 133-160.

Thorne, S.L., Black, R.W. & Sykes, J.M. (2009). Second language use, socialization, and learning in Internet interest communities and online gaming. *The Modern Language Journal*, 93, pp. 802-821.

Thorne, S. L., Fischer, I., & Lu, X. (2012). The semiotic ecology and linguistic complexity of an online game world. *ReCALL*, *24*(3), pp. 279-301.

Toyoda, E. & Harrison, R., 2002. Categorization of text chat communication between learners and native speakers of Japanese. *Language Learning & Technology*, *6*(1), pp.82–99.

Urh, M., Vukovic, G., Jereb, E., & Pintar, R. (2015). The model for introduction of gamification into E- learning in higher education. *Procedia - Social and Behavioral Sciences, 197*, 388-397. doi:10.1016/j.sbspro.2015.07.154

Van Lier, L. (2000). From input to affordance: Social-interactive learning from an ecological perspective. In J. Lantolf (Ed.), *Sociocultural theory and second language acquisition* (pp. 245-259). Oxford, England: Oxford University Press.

Van Lier, L. (2004). *The ecology and semiotics of language learning: A sociocultural perspective.* Boston: Kluwer Academic.

Lier, L. V. (2010). The ecology of language learning: Practice to theory, theory to practice. *Procedia - Social and Behavioral Sciences*, 3, 2–6. doi:10.1016/j.sbspro.2010.07.005

Voulgari, I., & Komis, V. (2011). Collaborative learning in massively multiplayer online games: A review of social, cognitive and motivational perspectives. In P. Felicia (Ed.), *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches*. Hershey, PA: IGI Global, pp. 370-394.

Wang, Y. F., Petrina, S., & Feng, F. (2017). Virtual immersive language learning and gaming environment. *British Journal of Educational Technology*, 48, pp. 431-450. http://doi.org/10.1111/bjet.12388

Warburton, S. (2009). Second Life in higher education: Assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. *British Journal of Educational Technology*, *40*(3), pp. 414-426.

Warschauer, M. (1995). The motivational aspects of using computers for writing and communication, *Symposium on Local & Global Electronic Networking in Foreign Language Learning & Research*, Honolulu.

Warschauer, M. (1996). Motivational aspects of using computers for writing and communication. *Telecollaboration in foreign language learning*, 29-46.

Wehner, A. K., Gump, A. W., & Downey, S. (2011). The effects of Second Life on the motivation of undergraduate students learning a foreign language. *Computer Assisted Language Learning*, *24*(3), pp. 277-289. http://doi.org/10.1080/09588221.2010.551757

Weisberg, S. (2014). *Applied linear regression (4th ed.)*. Hoboken, NJ: John Wiley & Sons, Inc.

Wendel, J. (1997). *Planning and Second Language Narrative Production*. Unpublished Doctoral Dissertation, Temple University, Japan

West, R. (1994). Needs analysis in language teaching. *Language Teaching*, 27(1), pp. 1-19.

White, L. (1987). Against Comprehensible Input: the Input Hypothesis and the Development of Second-language Competence. *Applied linguistics*, *8*(2), pp. 95-110.

Wiecha, J., Heyden, R., Sternthal, E., & Merialdi, M. (2010). Learning in a virtual world: experience with using second life for medical education. *Journal of Medical Internet Research*, *12*(1). https://doi.org/10.2196/jmir.1337

Wilcox, L. M., Allison, R. S., Elfassy, S., & Grelik, C. (2006). Personal space in virtual reality. *ACM Transactions on Applied Perception* (TAP), 3(4), pp. 412-428.

Wilks, D. (n.d.). *Playing in the MUD*. Available at: https://www.pcpowerplay.com.au/feature/playing-in-the-mud,390950 (Retreived 29 July 2015).

Willis, J. (1996). *A Framework for Task-Based Learning*. Longman, Harlow, UK.

Willis, D. & Willis, J. (2007). *Doing task-based teaching*. Oxford: Oxford University Press.

WoWWiki. (2016). *Quests.* Available at: http://www.wowwiki.com/Quest (Accessed: 12 October 2017).

Yanguas, I. (2010). Oral computer-mediated interaction between L2 learners: It's about time! *Language Learning & Technology*, *14*(3), pp. 72-93.

Yanguas, I. (2012). Task-based oral computer-mediated communication and L2 vocabulary acquisition. *CALICO Journal*, *29*(3), pp. 507-531. http://doi.org/10.11139/cj.29.3.507-531

Yuan, F., & Ellis, R. (2003). The Effects of Pre-Task Planning and On-Line Planning on Fluency, Complexity and Accuracy in L2 Monologic Oral Production. *Applied Linguistics*, *24*(1), pp. 1-27. doi:10.1093/applin/24.1.1

York, J. (2012) English Quest*. Modern English Teacher 21*(4), pp. 20-25.

York, J. (2014). Minecraft and language learning. In C, Gallagher (Ed.), *Minecraft in the Classroom: Ideas, inspiration, and student projects for teachers.* Berkeley, CA: Peachpit Press, pp. 179-196.

York, J., & DeHaan, J. W. (2017). Board games and foreign language learning: Rationale and framework development. *The PANSIG Journal 2016*, 379-390.

York, J., & deHaan, J. (2018). A constructivist approach to game-based language learning: Student perceptions in a beginner-level EFL context. *International Journal of Game-Based Learning*, *8*(1), pp. 19-40. http://doi.org/10.4018/IJGBL.2018010102

York J., deHaan J., Hourdequin P. (2019). It's Your Turn: EFL Teaching and Learning

with Tabletop Games. In: H. Reinders, S. Ryan, & S. Nakamura (eds) *Innovation in Language Teaching and Learning. New Language Learning and Teaching Environments*. (pp. 117-139). Palgrave Macmillan, Cham.

Yuksel, D. (2003). Same Task, Different Activities? A Replication of Coughlan and Duff's (1994) Study. *Proceedings of the Second International Online Conference on Second and Foreign Language Teaching and Research*, (1994), pp. 193-206.

Zalbidea, J. (2017). "One Task Fits All"? The Roles of Task Complexity, Modality, and Working Memory Capacity in L2 Performance. *The Modern Language Journal*, *101*(2), pp. 335-352. http://doi.org/10.1111/modl.12389

Zhao, Y., & Lai, C. (2009). MMORPGS and foreign language education. In R.E. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education*. New York, NY: IDEA, pp. 402-421.

Zheng, D. (2012). Caring in the dynamics of design and languaging: Exploring second language learning in 3D virtual spaces. *Language Sciences*, *34*(5), 543–558. https://doi.org/10.1016/j.langsci.2012.03.010

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System*, 56, 40–53.

Zheng, D. P. and Newgarden, K. (2012) Rethinking language learning: Virtual worlds as a catalyst for change. International Journal of Learning and Media, *3*(2): pp. 13–36.

Zichermann, G., & Cunningham, C. (2011). *Gamification by design: Implementing game mechanics in web and mobile apps*. O'Reilly Media, Inc.

Ziegler, N. (2016a). Synchronous Computer-Mediated Communication and Interaction. *Studies in Second Language Acquisition*, *38*(3), pp. 553-586. http://doi.org/10.1017/S027226311500025X

Ziegler, N. (2016b). Taking Technology to Task: Technology-Mediated TBLT, Performance, and Production. *Annual Review of Applied Linguistics*, *36*, pp. 136-163. http://doi.org/10.1017/S0267190516000039