Scale Validation Analyses
The Suicidality Scale Development Studies

Open & Sustainable Research – Psychometrics

Keith M. Harris, PhD

Charles Sturt University

Contents

This **Presentation**

- This presentation is intended for anyone interested in latent trait **scale development**
  - Scale development, research use, selecting scales, interpreting results
- Suicidality Scale development paper readers
  - Those interested in viewing an **Open Methods** discussion of more details on how we developed the SS
- Mental health professionals and students
  - Using scales to help form **clinical decisions**, diagnoses, etc.

- Intro > Background > Demonstrations > Interpretations > Applications

Overall **Aims**

- Provide info on **conducting** similar **studies** and using scales
- Provide **open methods** info for Suicidality Scale projects
- Contribute to **education** on scale development practices
- Encourage **localization** of scales by language/culture
- Encourage **sustainable** psychological science

- **Exchange** among professionals, students and community
  - Steps toward open and sustainable practices

Supporting Institutions & **People**

- **Chinese & English teams:** Keith M. Harris[1,2*], Lu Wang[3], Guanglun M. Mu[4,5], Yanxia Lu[6], Cheryl So[7], Wei Zhang[8], Jing Ma[9], Kefei Liu[10], Wei Wang[11], Melvyn W. Zhang[12], Roger C. Ho[13]
  - [1]School of Psychology, Charles Sturt University, Australia; [2]School of Psychology, University of Queensland, Australia; [3]School of Environmental and Life Sciences, University of Newcastle, Australia; [4]Education Futures, University of South Australia, Australia; [5]School of Teacher Education and Leadership, Queensland University of Technology, Australia; [6]Department of Medical Psychology and Ethics, School of Basic Medical Sciences, Shandong University, China; [7]Independent, Hong Kong; [8]School of Medicine and Health Management, Huazhong University of Science and Technology, China; [9]School of Politics and Public Administration, Zhengzhou University, China; [10]Yale School of Medicine, Yale University, USA; [11]Department of Psychology, Norwegian University of Science and Technology, Norway; [12]Biomedical Institute for Global Health Research and Technology, National University of Singapore, Singapore; [13]Department of Psychological Medicine, National University of Singapore, Singapore
- **Colombian Spanish team:** Castaño, M, Arenas, A, Pastrana, K, Van den Enden, P, Castro, J, Fandiño. O., Harris, KM
  - Universidad de Caldas, Charles Sturt University

The University of Queensland Australia

Charles Sturt University

Universidad de Caldas

Presenter

Keith M. Harris

- PhD – Psychology, The University of Queensland, Australia, 2009
  - Examinations of how suicidal people use the internet for suicide-related purposes
- MA – Social Psychology, Claremont Graduate University, USA
- BS – Psychology/Political Science, Michigan State University, USA

- Currently teaching psychopathology, postgraduate research methods
- Conducting research on mental health and information technologies, climate, scale development, and suicidality

- **NOT** a psychometrician!
  - For **expert guidance** in psychometrics, go to the real experts, see References for some
- Apply psychometrics to answer research questions, improve methods

UN Sustainable Development Goals

https://sdgs.un.org/goals

This project is driven to provide quality outputs that are free culture and localized. We aim to support the UN SDGs in the following ways

- **Good Health and Wellbeing** improving health and mental health assessments requires **good science**
  - We develop and advocate for **free** and **open** measures
  - Our instruments include collaborative development in local languages and cultures
- **Quality Education** through exchange of assessment, research and psychometric knowledge and skills, and by encouraging cultural diversity
- **Gender Equality** through collaborating and promoting women, girls and nonbinary^{+} leadership in research and clinical practice
- **Global Partnerships** by enhancing regional and international cooperation and access to science, innovation and knowledge sharing

Starting with the Classics

Classical Test Theory (**CTT**) & Scores

1a 1b 1c 1d 1e

1 = latent trait, the **construct** of interest; a, b, c.. are **attributes**, unique aspects of the same trait

Sum score = 1a + 1b + 1c..; Assumes **tau-equivalency** – all items are equal

See **References** for several papers on measurement models and the limitations of CTT

Terms (see References for further info)

- **Validity** – does the scale/instrument measure what it is supposed to measure?
- **Factor** – latent trait construct, items correlate 'load' on factors
  - E.g., suicidality, depression, extroverted personality, emotional intelligence
    - DASS-21 (Depression, Anxiety, Stress Scales) has three factors: depressive symptoms, anxiety symptoms, etc.
    - Suicidality scales (e.g., SABCS, SS) generally aim for one factor – suicidality/risk
  - **E**xploratory **F**actor **A**nalysis (EFA) – most common scale development analysis
- **Unidimensional** – only one factor, no subfactors
- **Loading** – individual item's association with factor (range -1.0 – 1.0)
  - Absolute higher value ≈ better, e.g., **-.75** is a stronger loading that **.68**
- **Communalities** (h^2) – degree other items explain one item (0 – 1.0), higher ≈ better
- **Fit** – degree scale fits the factor structure/model (range 0 – 1.0; near 1.0 is good)
  - Example: Tucker-Lewis Index (**TLI**), cluster fit
- **Error** – error fitting the factor structure/model (0 – 1.0, near 0 ≈ good)

More **Terms**

- **Theta** – the quantification of the latent trait, x-axis in IRT graphs
  - Typical range -4.0 – 4.0, mid-point near 0, end-points ≈ extreme levels of trait
- **Cutoff**, cut points – distinct scale scores differentiating low/med/high
  - The Hamilton Anxiety Scale (HAM-A): < 17 = mild, 18-24 moderate, > 24 severe
  - Area under the curve (AUC) – used to form cutoff scores
- **Tau-equivalent** – basis of CTT, all items have equal weighting etc.
  - Test requirement for Cronbach's alpha, AUC, CFA
- **Congeneric** model contrasts with tau-equivalent, **items vary** in weighting etc.
- Common **variance** – degree scale explains changes in construct scores
- **Reliability** – consistency between items, between assessments over time
- **DIF** – Differential item functioning, do scale items assess trait equivalently across groups (e.g., sex, age bands, first-language)

Measurement **Models** (see References for further info)

- **Hierarchical Cluster Modeling** – similar to EFA, how many clusters, loadings, includes **fit** and **error** (cluster model graph)
- **Factor analysis** – many types, do NOT use PCA; use ML, PAF, **min-res** (minimum residual, robust to skew)
  - psych package provides loadings, h^2, fit (TLI), error, common variance
- **Bifactor analysis** (EBFA, BA) – includes general factor loadings, h^2, ECV, error; general & group factor loadings, (bifactor graphic)
- **Item Response Theory (IRT)** – many models, test data first for which model
  - Graded response model (**GRM**) used for SS and many others
  - GRM includes item discrimination, information functions, cut points on theta, item details, scale coverage of theta
- Scale internal **consistency** – McDonald's **omega**
  - Suitable for congeneric models (items vary in weight etc.)
  - Close results to alpha, but **Cronbach's alpha** requires all items equal (tau-equivalent)

Aims & **Subjectivity**

- **Caveat**: while we may make suggestions, there will be many alternative approaches to achieving similar goals and different views on interpreting results

- Seminar **aims**: Introduce **newer**, **valuable** measurement **models**
  - EFA; EBFA; HCA; IRT (GRM; [DIF, scores]) – Brief Demonstrations

  - Our psychometric aim: What set of items best measures ‘suicidality’?

Suicidality Scale Studies

English 1, N = 5,115, English 2, N = 814, English 3, N = 626; All online surveys

Chinese Simplified, N = 1,595, Chinese Traditional, N = 1,393; All online surveys

Colombian Spanish, N = 313; Clinical and online survey, also clinician ratings

The **Suicidality Scale 1.0**

See the manual, link below, for more details

Code	Item/prompt	Responses
Ideation	How often have you thought about killing yourself in the past year?	1 = Never, 5 = Very often
Debate	In the past year, have you had an internal debate/argument (in your head) about whether to live or die?	1 = Never, 5 = Very often
Dead	Recently, have you been bothered by thoughts that you would be better off dead?	1 = Never, 5 = Very often
Meaning	Recently, have you felt your life is meaningless?	1 = Never, 5 = Very often
WTD	Recently, how much do you wish to die?	1 = Not at all, 5 = Very much
Predict	How likely is it that you will attempt suicide someday?	1 = Not at all, 5 = Very likely
RFD		1 = My reasons for living are greater than my reasons for dying, 5 = My reasons for dying are greater than my reasons for living = 5
DKS		I have no desire to kill myself = 1, I have a strong desire to kill myself
	Supplementary items *(not included in calculations)*	
WTL*	Recently, how much do you wish to live?	5 = Not at all, 1 = Very much
Attempt	Have you ever attempted to kill yourself?	1 = Never; 2 = Yes, but never really wanted to die, 3 = Yes, but was uncertain about dying, 4 = Yes, and at least once really wanted to die
Plan	Have you ever made a plan to kill yourself?	1 = Never, 2 = Yes, but never really wanted to die, 3 = Yes, but was uncertain about dying, 4 = Yes, and at least once really wanted to die

Publications & Open **Resources**

- You can find further details in manuscripts related to this talk
  - English Suicidality Scale development: Harris, K. M., Wang, L., Mu, G. M., Lu, Y., So, C., Zhang, W., Ma, J., Liu, K., Wang, W., Zhang, M. W., & Ho, R. C. (2022). Measuring the suicidal mind: The 'open source' Suicidality Scale, for adolescents and adults. *Preprint*. https://doi.org/10.31219/osf.io/b4qut
  - Colombian Spanish Suicidality Scale: *In preparation*.
  - Chinese Suicidality Scale development: Wang, L., Harris, K. M., et al. (2022). Improving Chinese suicide risk assessment: Development of the Chinese Suicidality Scale. *In preparation*.
- Open data and resources
  - English study data: https://osf.io/vjxnq/
  - English SS manual: https://osf.io/6tknd/

Developing a Scale

Main **Steps** in Scale Development/Validation

See more detailed guidelines in the **References** and elsewhere

- This is just a very brief presentation of our steps

- Start with a **quality item** set, include an **item pool** when possible
  - *Example*: Colombia study used 8 validated SS items, plus 2 items
- **Community testing** – test the language/meanings with small community samples
- **Open Discussion** – project team discuss wording, community responses, other scales, items
- **Collect good data** – consider the *population* (community, clinical)
  - You want to **cover the spectrum** of your variables (suicidality, low to high)
  - Collect **enough data**, as many participants as possible

Scale Development – **First** Steps

- Choose **quality** items
  - Previously published scales, examine EFA and other statistics to choose best
  - Expert recommendations
  - Personal experience (critically examine your own data)
- Item **Pool** – bigger is better, sort of. Too many items will fatigue participants, remove poor and redundant items
- Conduct sound **survey methods** (see References)
- **Cleanse** data and **replace** missing values (sources in References)

Community Sampling

- After determining near-final versions of our scale, item pool
  - We sent brief questionnaires to community locations
  - We included, one by one, the Scale instructions, and each item
  - We asked community members to **rate** each item: Low, Medium, High
    - **Clarity** – is this easy to understand? Is anything unclear?
    - **Validity** – could you answer this questions accurately, as worded here?
    - **Suggestions** – we asked people to provide suggestions on improving the wording
  - Next, the team evaluated community responses and finalized the items

- Next slide, examples from the Colombian-Spanish and Chinese (simplified and traditional scripts) projects

Community Sample Results

Colombian-Spanish

T1P = *De 1 a 6 que tantas razones tiene para vivir?, siendo 1 que sus razones para vivir superan las razones para morir; y 6 que sus razones para morir superan sus razones para vivir.*

T1R1 = *(1)Mis razones para VIVIR son mayores que mis razones para morir.*

1. **"Este énfasis con mayúsculas me parece importante".**
2. **"Consideraria modificar esta opción de la siguiente manera "Tengo más razones para vivir que para morir" de esa manera me parece mas claro".**

T1R6 = *(6) Mis razones para MORIR son mayores que mis razones para vivir.*

Chinese

Simplified	Clarity	Validity	Comments
以下问题是关于自杀的私人问题。请选择最适合您的选项。	M	M	"个人问题" 比 "私人问题"更好
Traditional	**Clarity**	**Validity**	**Comments**
我們需要問你一些有關自殺的個人問題。請標出與你本人情況最相符的選項。	M	H	1."我們需要問你一些有關自殺個人問題" could be changed to"以下是與自殺有關的私人問題"。3."標出"could be changed to"選出"

Community Sampling **Outcomes**

Why do community sampling?

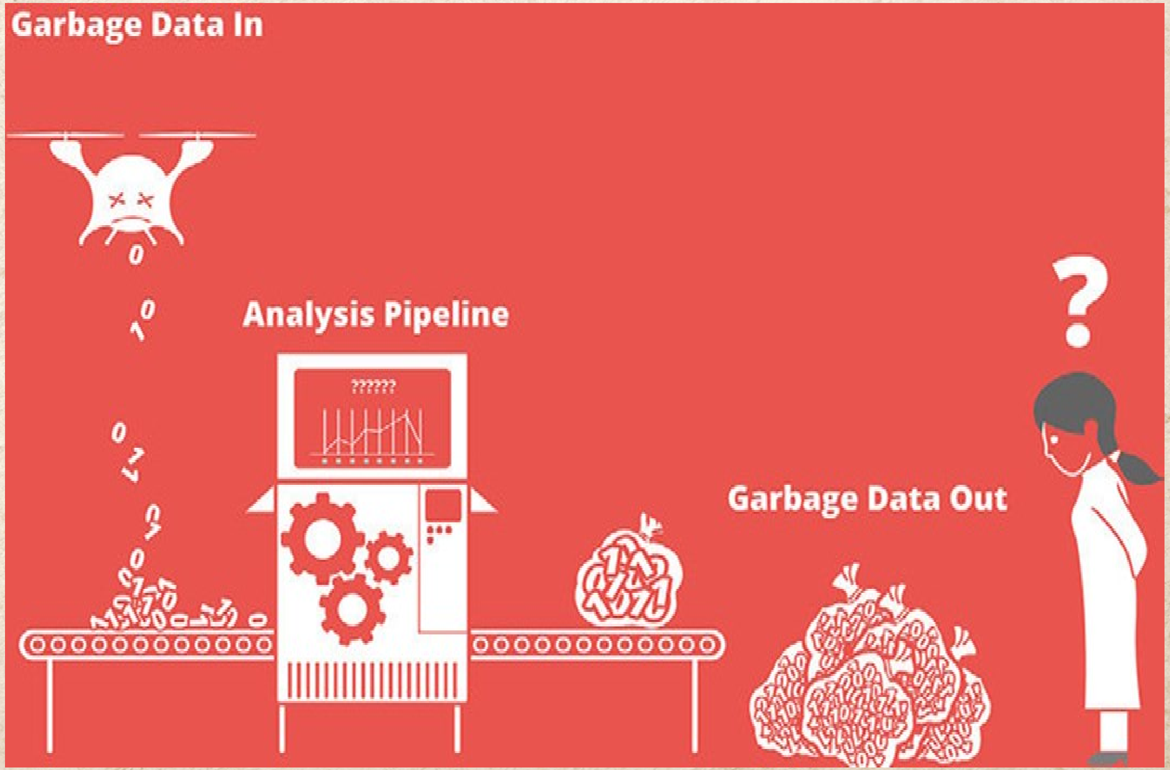
- We found some items, could be made clearer, including in English
  - E.g., For reasons for dying (RFD), the original item stated "my reasons for dying **outweigh** my reasons for living." For younger people and non-native speakers, we thought "are greater than" worked better than "outweigh"
  - Usually, we want to make scales and items as concise as possible, but **validity** outweighs nearly all else
- For other languages (e.g., Colombian-Spanish, Chinese traditional script), there were quite different ways of expressing a specific cognition or affect. Many items underwent several revisions and further testing
- This also fits with our **SDG** goals for localizing these instruments and giving more voice to community members

In **Sum**

To measure traits accurately and produce valid findings…

Avoid Garbage In – Garbage Out **GIGO**

- Cleanse dataset, replace missing values, etc. (other seminars)
- Validate **Factors/Dimensions**
- Validate **Items**
- Finalize Best-possible **Scales**
- **Report** Scale Diagnostics



We Have Good Data, What **Next**?

Data

- Checking the Data – how can we **use** our data?
  - Check **frequencies** for all variables
  - Are any variable response options **missing**?
  - Example: Item scored 1 – 7, but no responses for ‘6’
  - Any **odd** patterns?
  - Example: Only a few responses at low end, many high

  - **See** your data – Rest-score plots, Histograms

SS Examples - **Frequencies**

These are basic descriptive statistics of two pool items. Our main concern here is that each response option is endorsed.

		AttemptB	AttemptIntent
N	Valid	5115	5115
	Missing	0	0
Mean		.29	2.15
Std. Deviation		.454	1.556
Skewness		.919	.953
Std. Error of Skewness		.034	.034
Kurtosis		-1.156	-.749
Std. Error of Kurtosis		.068	.068

AttemptB = binary item, have you attempted suicide? No = 0, Yes = 1
Attemptintent
Have you attempted suicide?
1 = never, 2 = yes, but didn't really want to die, 5 = yes and really wanted to die

AttemptB

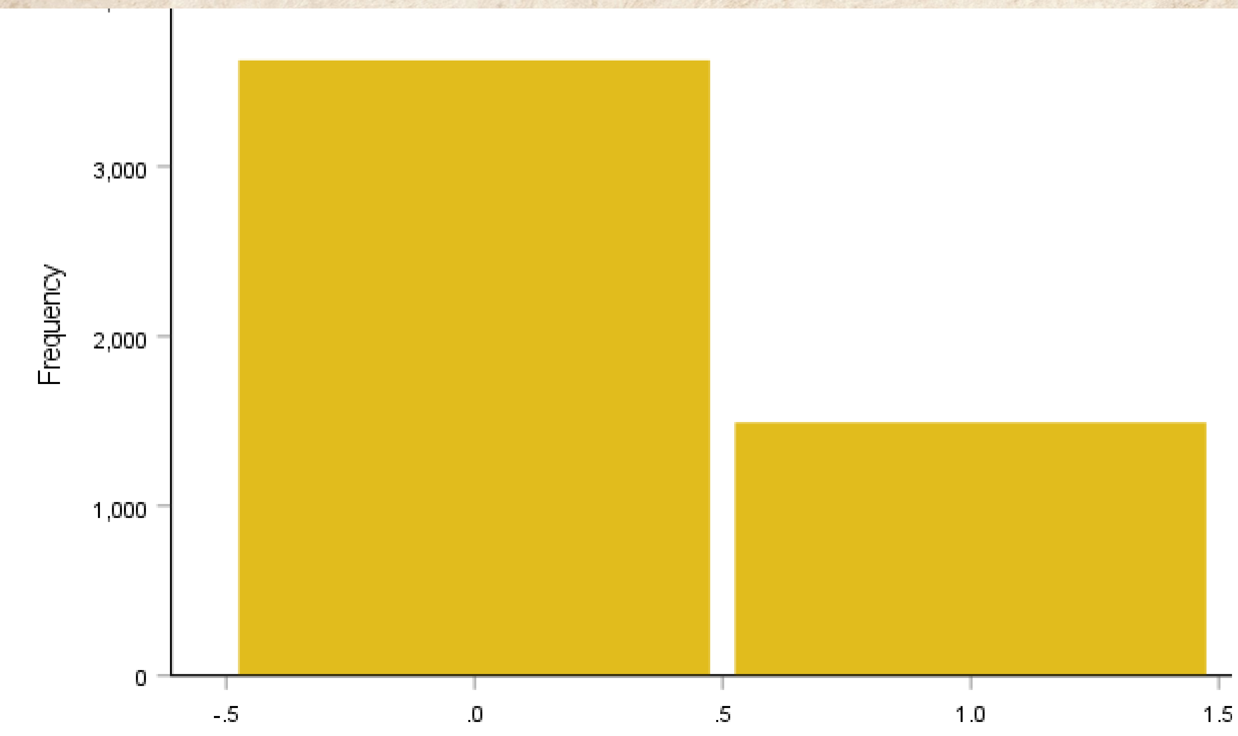
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3625	70.9	70.9	70.9
	1	1490	29.1	29.1	100.0
	Total	5115	100.0	100.0	

AttemptIntent

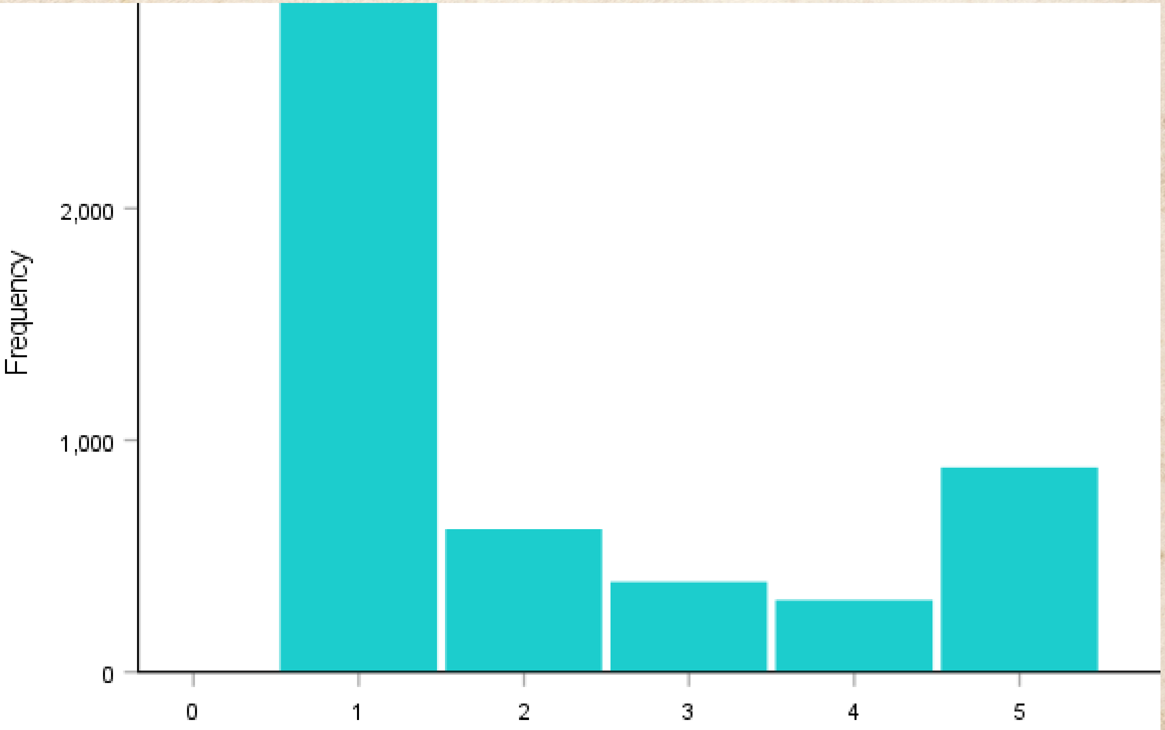
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2915	57.0	57.0	57.0
	2	616	12.0	12.0	69.0
	3	390	7.6	7.6	76.7
	4	312	6.1	6.1	82.8
	5	882	17.2	17.2	100.0
	Total	5115	100.0	100.0	

Notice that many more reported 'no suicide attempt' with the binary version than the continuous version. **Why?**

Item Response **Frequencies**



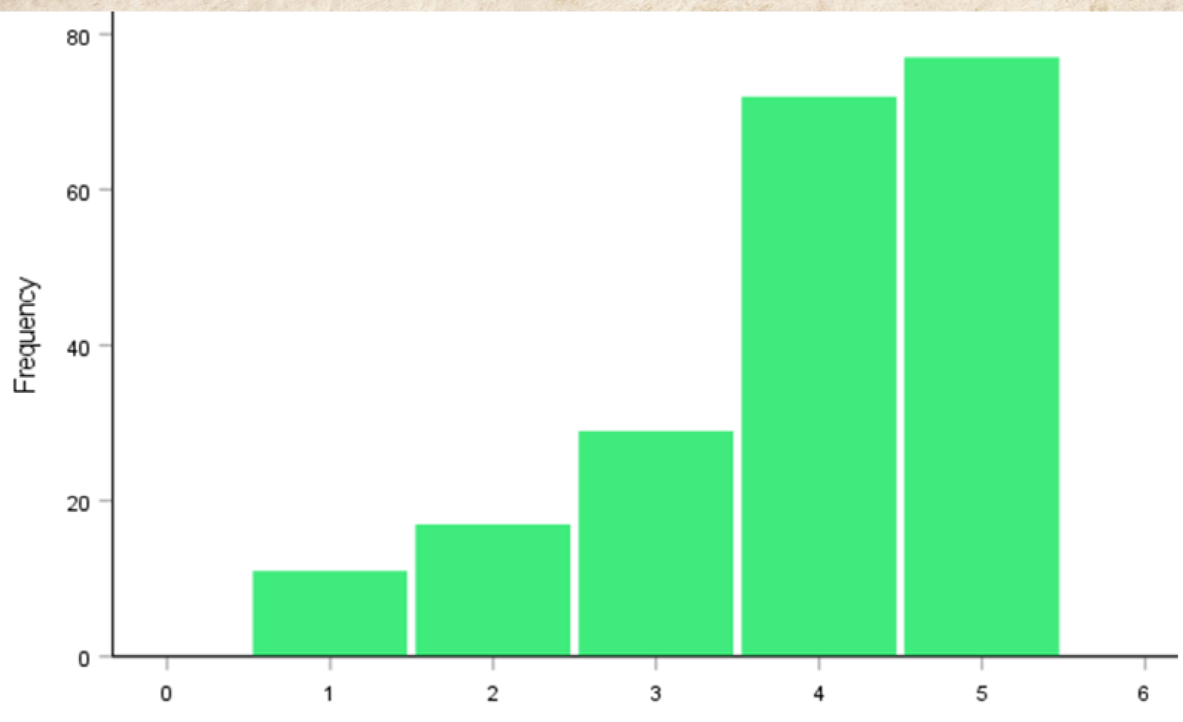
Attempt **Binary**

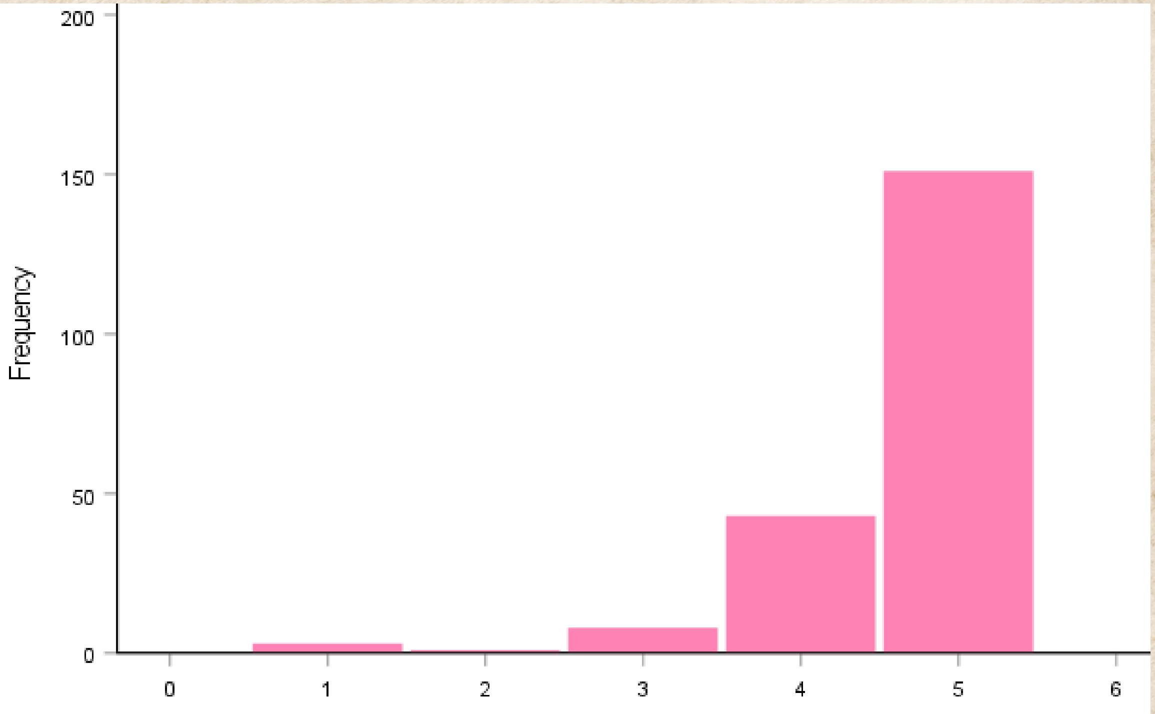


Attempt with **Intent** to die

These histograms show the frequencies of the items from the previous slide. We are looking for any responses that are under-endorsed. We are not really concerned by skew here. Which item do you think would provide more information on suicidality/risk?

Other Examples: **Item** *or* **Sample?**

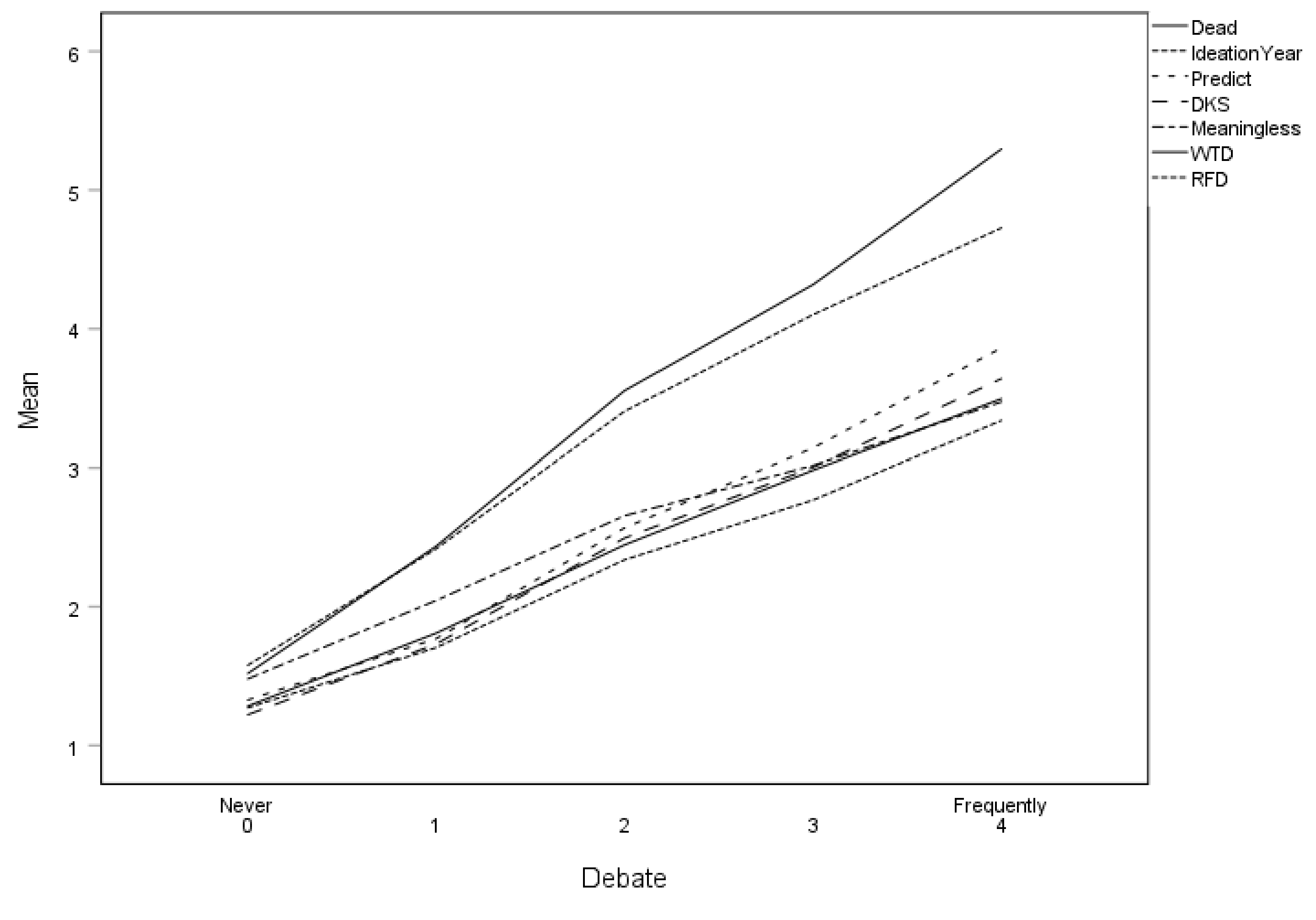




The item on the left appears fine, skew is not usually a problem at the item level

The item on the right has serious problems.
What would you do with this? Rescore, delete, something else?
Next, is the problem the **item** (poorly formed?), or the **sample** (not diverse?)**?**

SS – **Rest-Score Plot**: Debate

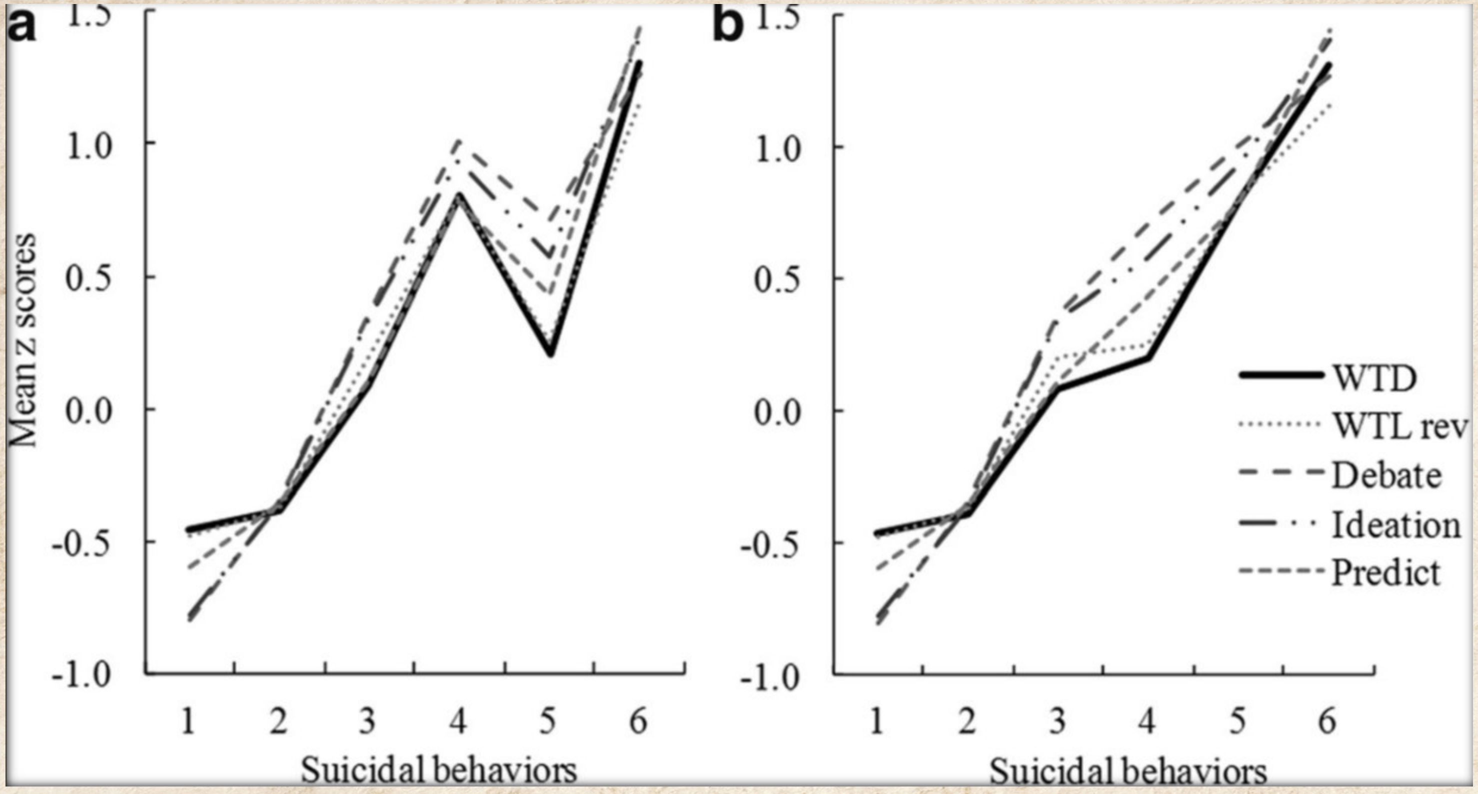


This is a rest-score plot. We put one item on the x-axis, then examine how it correlates with other items.

This works best when we have some good/solid items to compare the reference item to.

Note that there is a (reasonably) consistent linear relationship between all variables

Rescore?



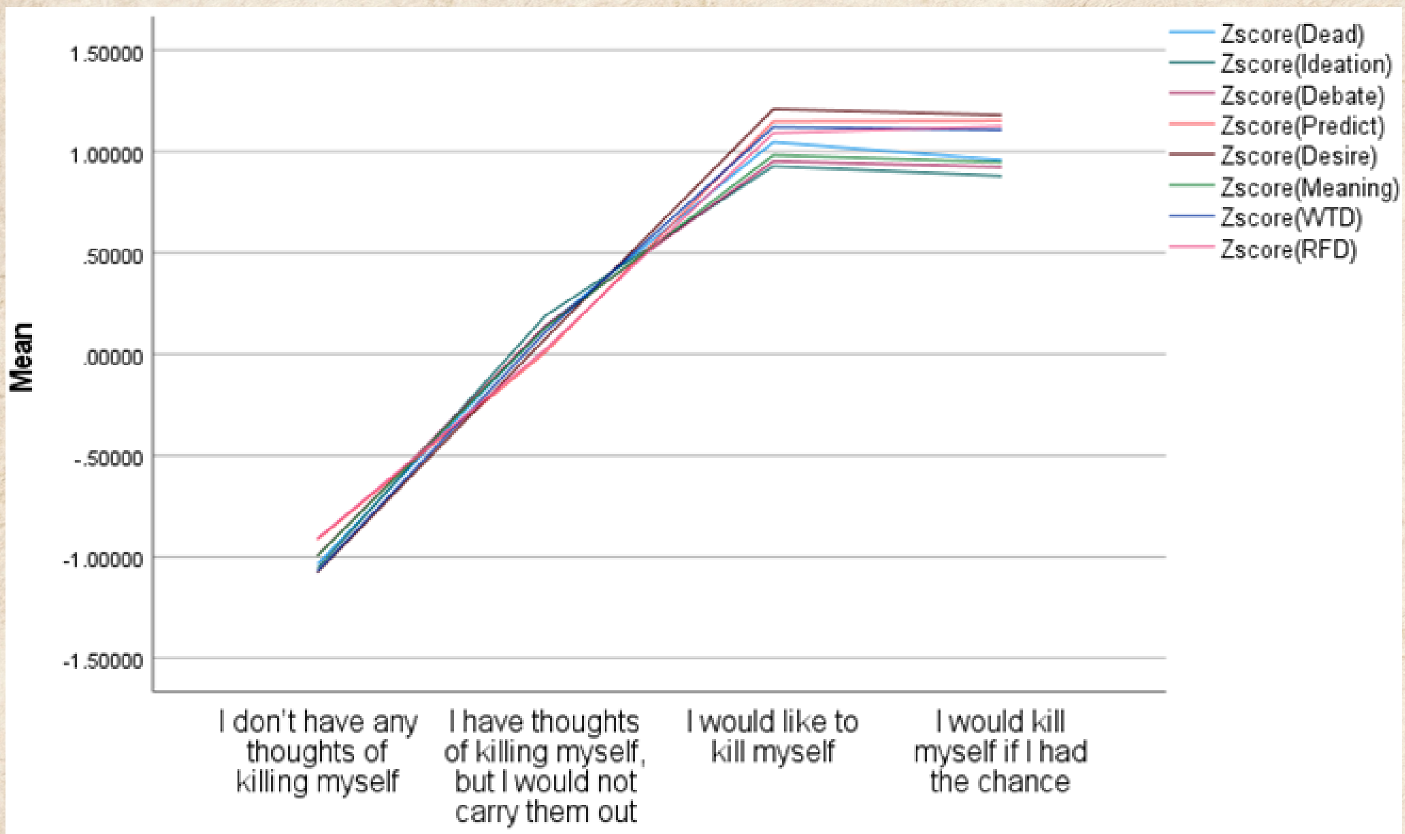
Source: Harris, K. M., Lello, O. D., & Willcox, C. H. (2017). Reevaluating suicidal behaviors: Comparing assessment methods to improve risk evaluations. *Journal of Psychopathology and Behavioral Assessment, 39(1), 128-139.*

Guttman item – ordered categories. Google for examples.

Monotonicity – consistent order, item responses demonstrate consistent increasing or decreasing levels of the latent trait

The graph on the left shows the original scoring of a Guttman-type item on suicidal behaviors as it correlates with 5 items from the Suicidal Affect-Behaviors-Cognition Scale. Due to clear violations of monotonicity (in multiple studies and subsamples), it was rescored: 5 as 4, and 4 as 5, resulting in the figure on the right. Would you use the original item scoring shown on the left**?**

Item **Detective** Work

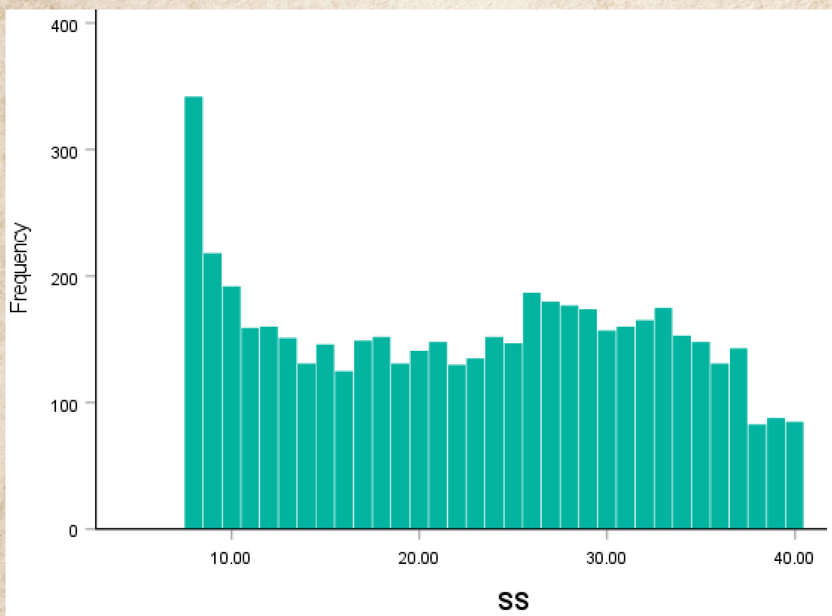


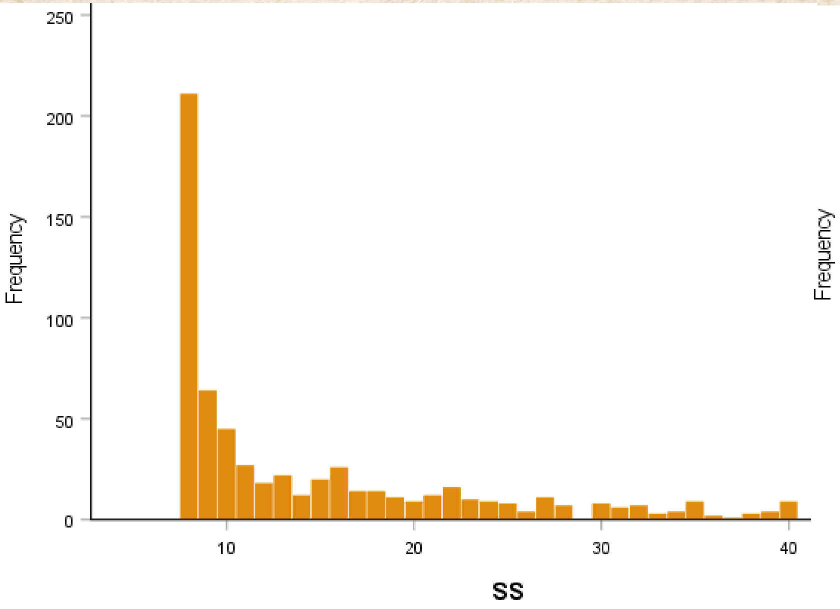
This rest-score plat shows a strong linear relationship over the first three of these **Guttman-type** responses. However, there is no difference between responses 3 and 4. Thus, **violating monotonicity**.

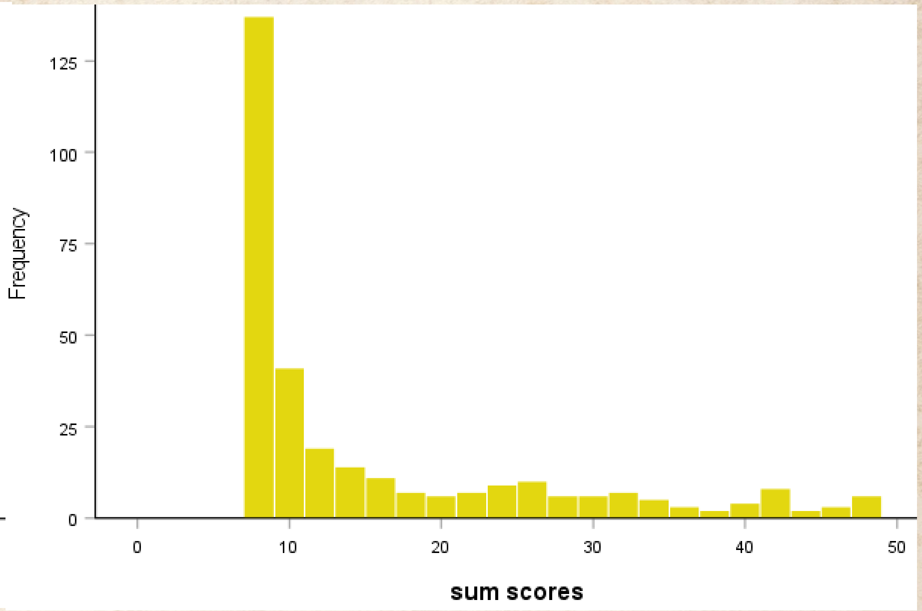
But, which is better, the 3rd or 4th option? Further research is needed to understand which, if either, might work best.

How might you design a study to test this?

Sample Construct **Coverage**



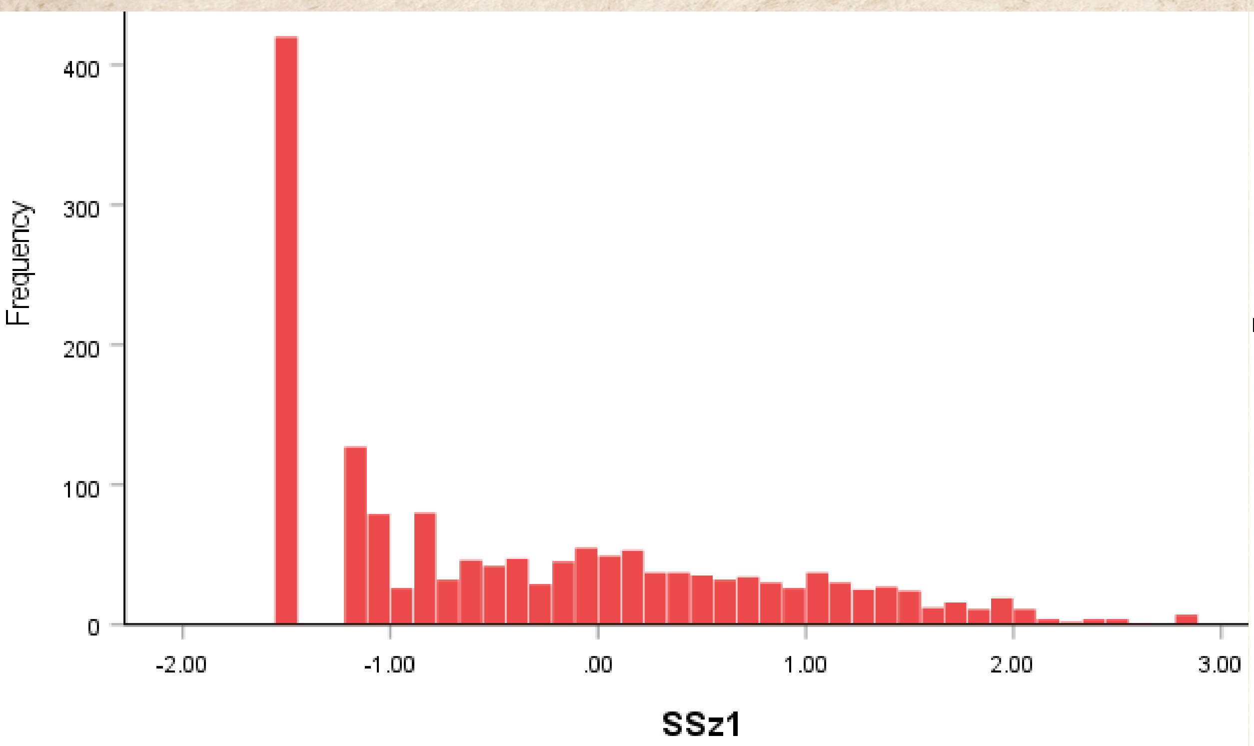


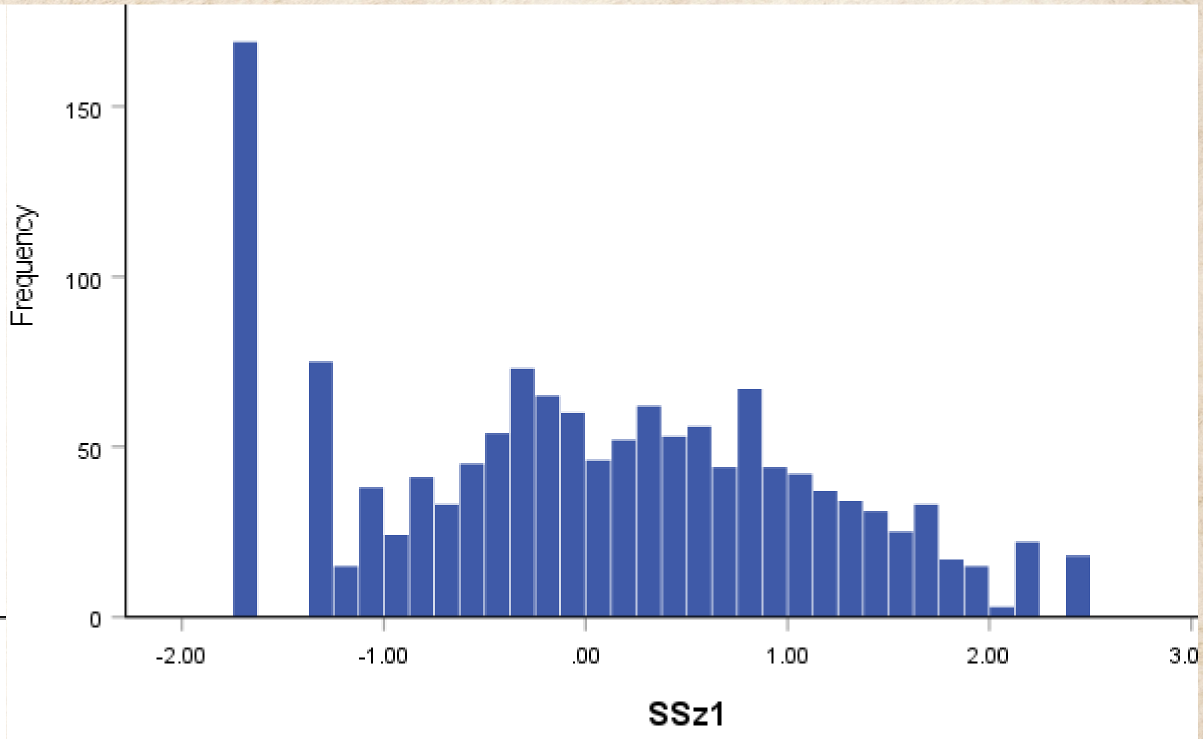


Here, we see three histograms of SS sum scores, from independent studies (English S1, S2, Colombia). Note that the left graph shows very good coverage of the full assessed range of suicidality. However, the other two graphs show small numbers of participants at high ranges. While common, to provide more definitive findings on scale properties, larger more diverse samples are needed.

How would you obtain a sample with large numbers of suicidal participants?

Coverage – Chinese (ability scores)





Simplified **Traditional**

Similar to the previous slide, these graphs show the sample range on suicidality, but through **IRT-derived ability scores**, which better approximate true suicidality levels (compared with sum scores).

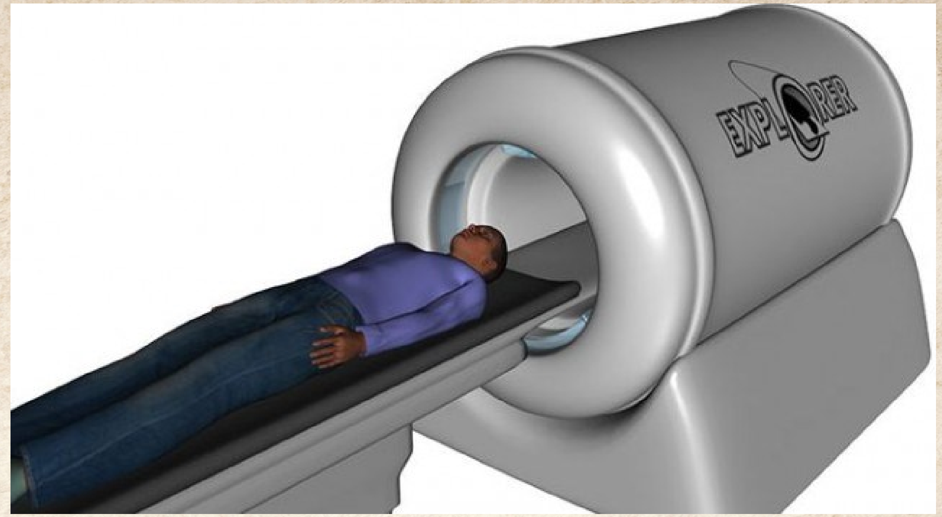
Choosing a Measurement Model

The **Model** & Your **Aims**

- Project aims: Best possible measure of suicidality
  - Strong item properties (linear relationships, monotonic, high discrimination, high loadings)
  - Strong scale properties (high **unidimensional** fit, low error, high internal consistency, high test-retest)
  - Consistency across demographic groups – age, gender, first-language, etc.

- Next, match your data with your measurement model
  - All known suicide risk assessment studies have shown suicide-related items fit a congeneric model (items have different weighting, etc.). No known study has validated the use of tau-equivalent models for this construct
- There are several measurement models to choose from (check References)
- Would you choose a toaster to make coffee? Would you choose a vintage X-ray machine to do a brain scan**?**

Measurement **Models**



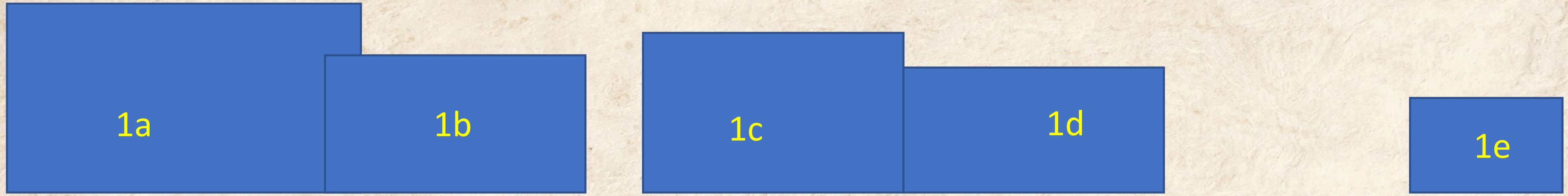
CTT Model Statistics

- These are the **most common** scale development analyses
- Green indicates flexibility in data/model
- Red stats require tau-equivalency (all items equal)
- Standards (e.g., > .70) are common reference points

- **EFA** to identify factors, determine *appropriate* items (loadings ≥ .32?)
- Cronbach's **alpha** for reliability, scale *quality?* (α ≥ .70?)
- **CFA** to validate *factors* in new samples (fit ≥ .90?, RMSEA < .08?)
- Pearson's ***r*** for test-retest reliability (*r* ≥ .70?)

Single or **Multi**model?

Are All Items Equal?



Factor analysis statistics (loadings, communalities, inter-item correlations), indicate varying associations between items/attributes. No known study has shown that all items of a latent trait scale are equal across psychometric tests.

Sum score = ?

An alternative scoring method uses *response patterns*, resulting in '**ability**' (aka factor/person/individual) scores. – see **IRT**

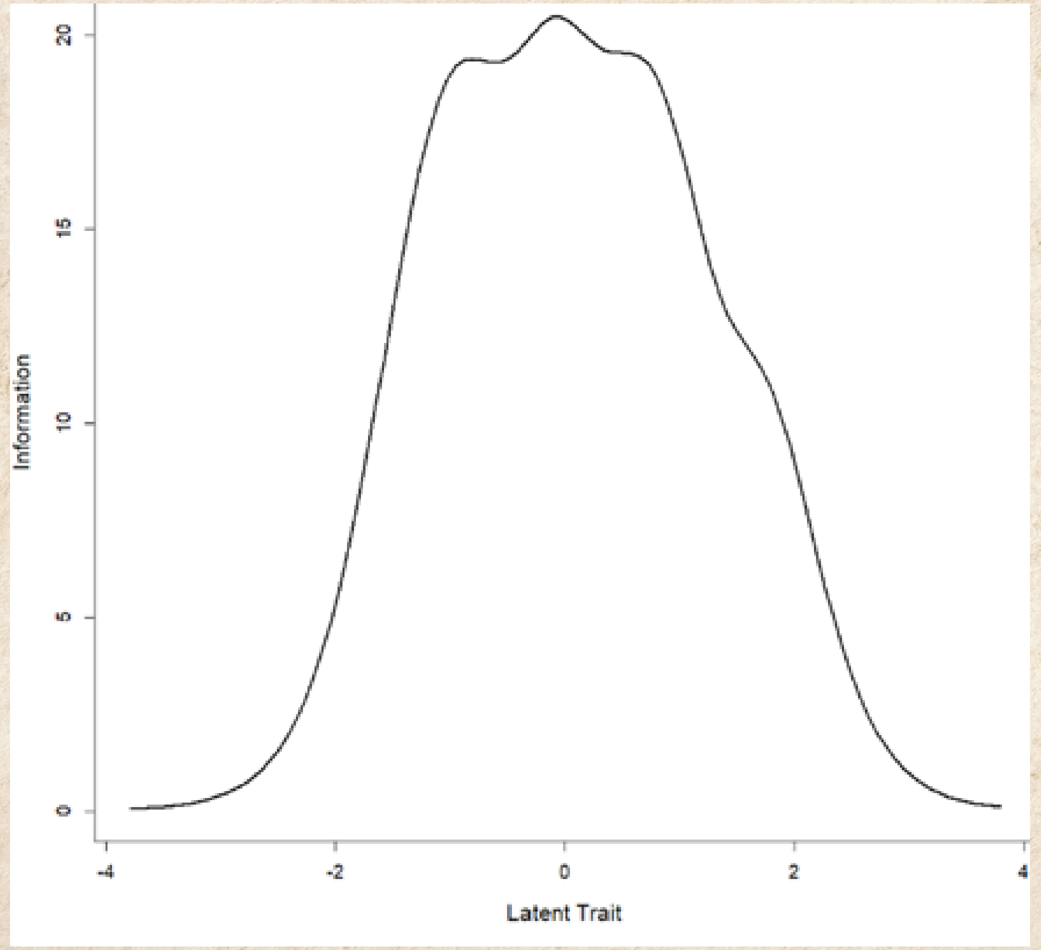
Multimodel **Scale Testing**

- Next, we use four measurement models to examine strengths and weaknesses in candidate items for a suicidality scale
- For the first three, HCA, EFA, BFA, we are looking for:
  - High item loadings (roughly .70 - .99)
  - High item communalities/h2 (roughly .60 - .99)
  - High model fit and unidimensionality (varies; .70 - .99)
  - High explained variance of the latent trait (roughly .70+)
  - Low model error (roughly .00 - .20)
  - Item and scale values can be compared with other included scales

Adding **IRT** to Multimodel Testing

- For IRT, we need to choose from various models
- We chose the graded response model (GRM)
  - GRM is flexible, allowing items to vary on various parameters
  - Within GRM, we need to determine if a constrained or unconstrained model fits best
  - Unconstrained – items can vary in discrimination (*a*)
- For IRT results, we look for
  - High item discrimination levels (*a*, roughly > 1.80)
  - Item response monotonicity/validity
  - Scales that cover a broad range of theta (the latent trait)

Suicidality Scale **Test Information** (IRT)



This line shows the SS **test information curve.** The area under the line indicates the **volume** of information the test provides on the latent trait. The location of the line shows **where** on theta the information is captured.
Note that, for all scales, less information – and less certainty, is found at extreme levels.

X-axis = Theta (latent trait)

Additional Item/Scale **Tests**

- Through GRM, we can test whether items or scales show evidence of **d**ifferential **f**unctioning, or fail invariance, across groups
  - **DIF**/DTF can test whether **i**tems/**t**ests measure the latent trait differently for females/males/nonbinary+; urban/rural residence; first-language, etc.
  - If items/scales fail these tests, scale scores can have qualitatively different meanings across groups
- We used McDonald's ω for testing scale internal consistency
  - Omega is suitable for congeneric models, while alpha requires tau-equivalency
  - Bootstrapping helps address subjectivity and sample-specific findings
- Please see References and other sources for more info

For most advanced stats we use the **open-source** R statistical environment

- R Core Team. (2022). *R: A language and environment for statistical computing*. In (Version 4.2.0) R Foundation for Statistical Computing. https://www.R-project.org/

We used these R packages

- **psych** Revelle, W. (2021). *psych: Procedures for psychological, psychometric, and personality research* [R package version 3.6.2]. Northwestern University. https://personality-project.org/r/psych/psych-manual.pdf
- **coefficientalpha** Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement*, *76*(3), 387-411. https://doi.org/10.1177/0013164415594658
- **ltm** Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1-25. http://www.jstatsoft.org/v17/i05/
- **lordif** Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1-30. http://www.jstatsoft.org/v39/i08/

R Basics

- **Download** latest version of R for Windows/Mac/Linux: https://cran.r-project.org/bin/windows/base/
- Download RStudio https://www.rstudio.com/products/rstudio/download/
- Get help: https://rstudio-education.github.io/hopr/starting.html

- Get latest **manuals** for packages
  - R **basics**: https://cran.r-project.org/manuals.html
  - Some **recommended** packages: https://support.rstudio.com/hc/en-us/articles/201057987-Quick-list-of-useful-R-packages
  - **Google** package name, e.g., “r package psych” to get link
    - https://cran.r-project.org/web/packages/psych/index.html

Getting Started in **R**

- **Define a scale/item** set (matrix)
- In R/Rstudio, define item set/scale, example: “scaleA”
  - Text within “quote marks” can be copied and pasted into R
  - “scaleA <- study1[, 3:10]”
    - This assigns a name ‘scaleA’ to a matrix
    - study1 = name of dataset; 3:10 = columns for scaleA variables in dataset
  - “scaleB <- study1[c(2,5,7,9,10)]” Use this method for noncontinuous columns

HCA Demonstration

- **psych** package (download and activate)
  - Extra **help** for psych: https://personality-project.org/r/psych/

- Type “iclust” to do HCA, and identify the item set/matrix (e.g., ‘scaleA’)
- “iclust(scaleA)”
- Copy and interpret results, see next slide

HCA Output (English S1 Item Pool)

Example R output

ICLUST (Item Cluster Analysis)
Call: iclust(r.mat = pool)
Purified Alpha: [1] 0.97
G6* reliability: [1] 1
Original Beta: [1] 0.68
Cluster size: [1] 30
Item by Cluster Structure matrix: [,1]

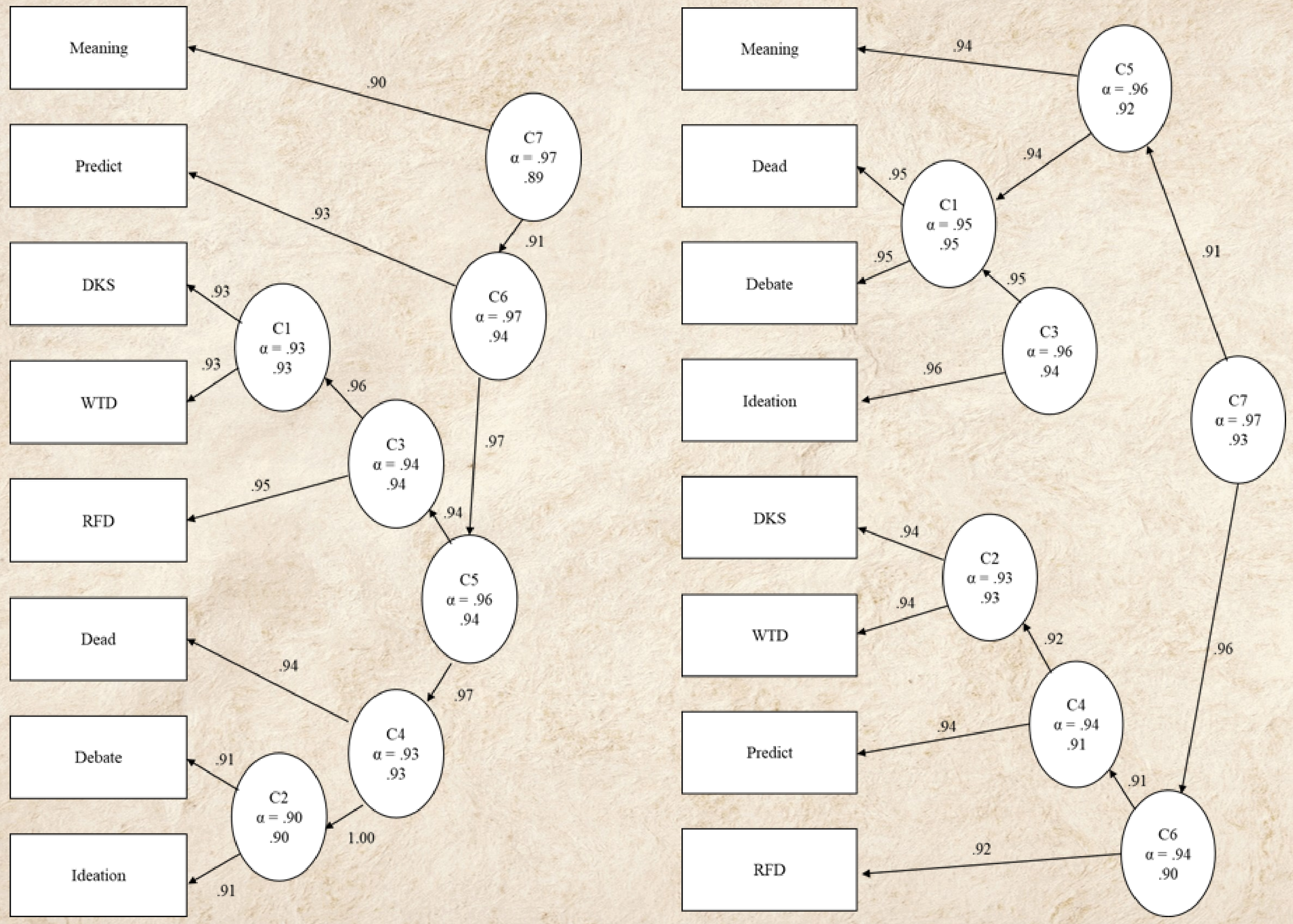
Dead	0.80
Ideation	0.83
StopThoughts	0.74
Deterrents	0.55
Reasons	0.65
Wish-HAMD	0.65

Cluster fit = 0.96 Pattern fit = 0.98 RMSR = 0.07

This is a selection of Item Pool output
Of note, in blue, you can see cluster 'loadings.'
In yellow, we see the cluster fit and error.
You can also see that only 1 cluster was identified [1], if there were 2, there would be loadings etc. for the second.

Main points to note are:
Number of clusters, strength of loadings, fit, error.

Suicidality Scale **Hierarchical Cluster Analyses** English S1 (left) S3 (right)



These are graphs from HCA results. They provide item cluster loadings, and info on how items relate to each other, and possible hierarchies of trait attributes.

EFA Demonstration

- psych package
- Define item set/scale, example: “scaleA”

- Choose FA type, ML, PAF, minres – we chose minres
- “mr <- fa(**scaleA**, fm = "minres", alpha = TRUE, values = TRUE)”
- “mr”
- Copy and interpret results

EFA Output

mr <- fa(pool, fm = "minres", cor = "mixed", alpha = TRUE, values = TRUE) [command, type this, use "cor = "mixed"" if you have both dichotomous and polytomous items, otherwise use default for polytomous]

> mr [type this]

Factor Analysis using method = minres

Call: fa(r = pool, fm = "minres", alpha = TRUE, cor = "mixed")

Standardized loadings (pattern matrix) based upon correlation matrix

	MR1	h2	u2	com
Dead	0.84	0.71	0.29	1
Ideation	0.87	0.76	0.24	1
Debate	0.89	0.79	0.21	1
SelfHarm	0.65	0.43	0.57	1
Stop	0.73	0.53	0.47	1

MR1 = factor 1, h2 = communality

Note loading, h2, other info useful too

Proportion Var 0.66

Mean item complexity = 1

Test of the hypothesis that 1 factor is sufficient.

Tucker Lewis Index of factoring reliability = 0.469

RMSEA index = 0.271

Note common variance indicates 1 factor is sufficient (unidimensional)

Note TLI (model fit)

Note RMSEA (error)

BFA Demonstration

- psych package
- Define item set/scale, example: “scaleA”

- “omega(**scaleA**)”
- Copy and interpret results

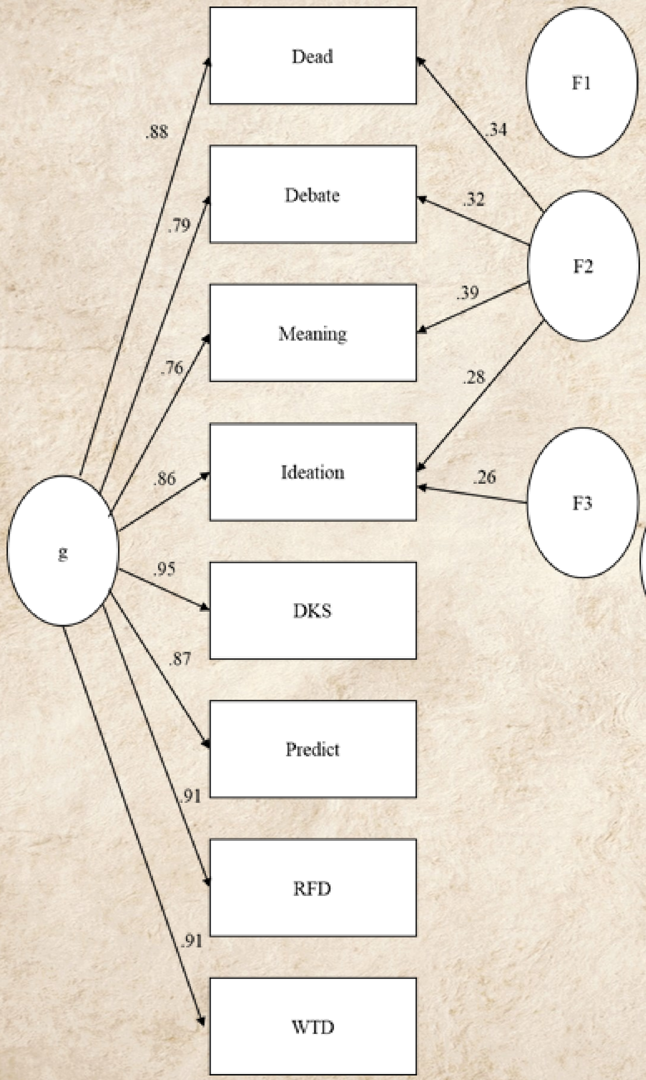
BFA Output

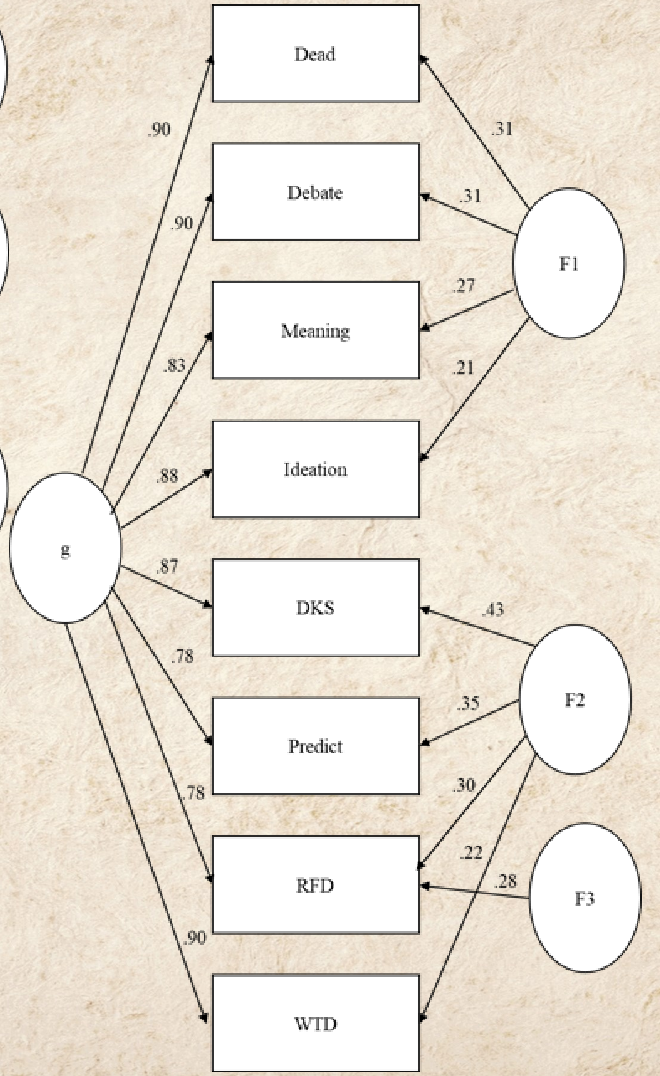
```
> omega(pool)
Omega
Alpha:                   0.97
G.6:                     0.98
Omega Hierarchical:      0.83
Omega H asymptotic:      0.85
Omega Total              0.98

Schmid Leiman Factor loadings greater than  0.2
          g      F1*  F2* F3*  h2
Dead      0.76   0.37          0.71
Ideation  0.78   0.27     0.23 0.73
Debate    0.79   0.26          0.73

Explained Common Variance of the general factor =  0.72
RMSEA index =  0.055
```

There are many statistics from BFA. Note the Omega-H; the g loadings, the h2
Also, examine the F1-F3 loadings

Note the ECV and RMSEA (error)

Suicidality Scale **Bifactor Analyses** English S2 (left) S3 (right)





Similar to the HCA diagram, the bifactor diagram shows **how some items relate** to others, and the overall construct.

g = general factor – the latent trait
F = group factor, item subgroups
Numbers = loadings, higher ≈ stronger

These relationships can help us understand how the latent trait works. Possibly opening the door to more sophisticated assessment of trait characteristics.

Here we see preliminary evidence that there may be two sets of four items, that show assessment strengths at low or high levels of the trait. How would you investigate that?

IRT – **GRM** Demonstration

- ltm package
- Decide/test appropriate model, GPCM, GRM, etc. We selected GRM
  1. Run constrained GRM model (all items have equal difficulty/value) = Fit1
  2. Run unconstrained GRM model (items are allowed to vary in difficulty/value) = Fit2
  3. Run ANOVA to test for information loss levels between Fit 1 and 2, lower AIC (information loss) is better, a significant p value means the AIC/BIC significantly differ, meaning the model/fit with lower AIC and BIC is better
  4. Run analyses with the better GRM fit
  - “`fit1 <- grm(data = scaleA, constrained = TRUE)`”
  - “`fit2 <- grm(scaleA, , IRT.param = TRUE, constrained = FALSE)`”
  - “`anova(fit1, fit2)`”

IRT – GRM – item statistics (*b*, *a*)

- ltm package

- "fit2 <- grm(**scaleA**, , IRT.param = TRUE, constrained = FALSE)"
- "fit2"
- "summary(fit2)"
- Copy and interpret results

GRM Output

```
grm(data = pool, constrained = FALSE, IRT.param = TRUE)
Coefficients:
$Dead
Extrmt1  Extrmt2  Extrmt3  Dscrmn
 -0.821   0.270    0.922    2.400

$Ideation
Extrmt1  Extrmt2  Extrmt3  Extrmt4  Dscrmn
 -1.504  -0.537    0.152    0.692    2.727
```

Note the discrimination values = *a*, the Extrmt values = b_1, b_2, …

See the next slide for an example

IRT – GRM (test & item **information functions**)

- ltm package

- Test/Item Information
- “information(fit2, c(-4,4))” full test (scale) information
- “information(fit2, c(-4,4), items = c(1))” item 1 (first) info
- “information(fit2, c(-4,4), items = c(2))” item 2 (second) info
  - Repeat for all items

Item Information Functions

information(fit2, c(-4,4), items = c(1))
Call:
grm(data = ss, constrained = FALSE, IRT.param = TRUE)
Ideation
Total Information = 11.55
Information in (-4, 4) = 11.55 (99.96%)
Based on items 1

> information(fit2, c(-4,4), items = c(2))
Call:
grm(data = ss, constrained = FALSE, IRT.param = TRUE)
Debate
Total Information = 9.95
Information in (-4, 4) = 9.9 (99.47%)
Based on items 2

Here are two SS items.
You need to type in the item name, in yellow, as this is not provided in R output. The only number we use here is the **Total Information** (**IF**) which we can put in our table. This tells us how much info on theta each item captures. We can compare this to the **Test Information** value to see proportion per item.

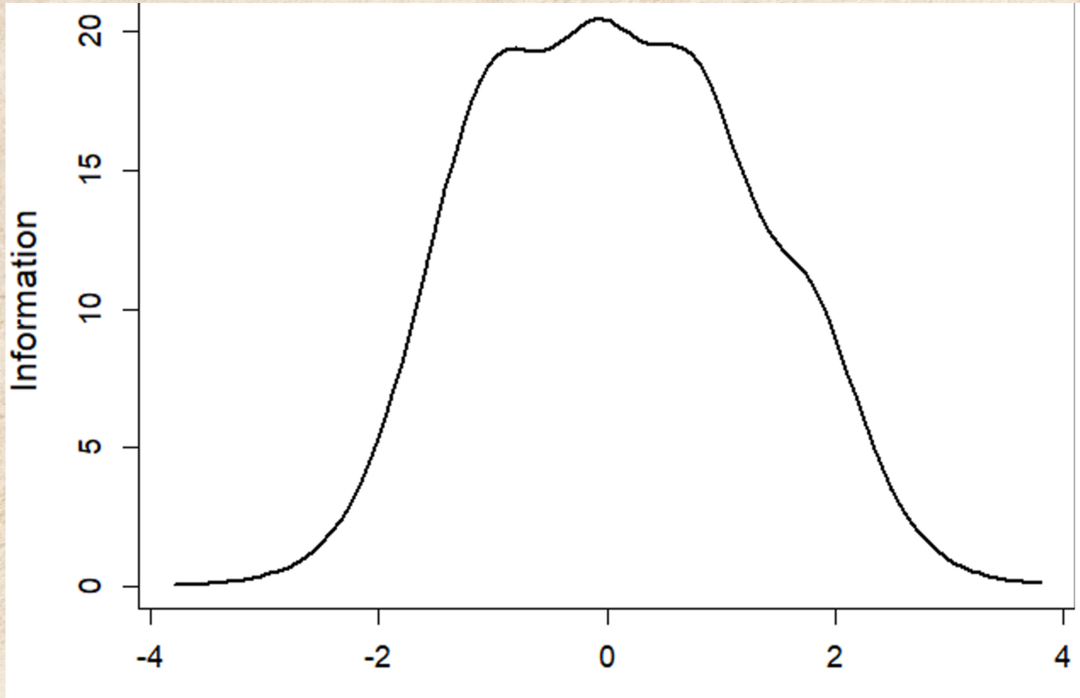
IRT – GRM - **IICs**

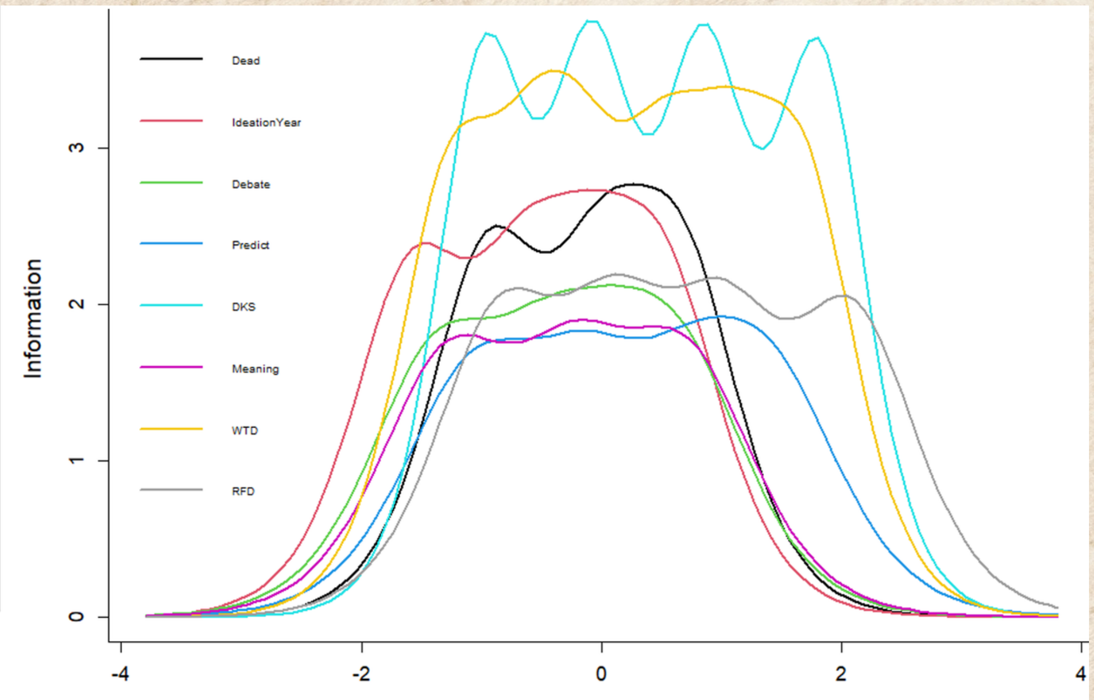
- ltm package

- Test/Item Information Curves
- “plot(fit2, legend = TRUE, type = "IIC", cex = 1.0, lwd = 2, cx = "topleft", xlab = "Latent Trait", cex.main = 1.5, cex.lab = 1.3, cex.axis = 1.1)”
- = all test items plotted together, compare info levels, range

- Item Response Category Characteristic Curves
- plot(fit2, lwd = 2, cex = 1.0, legend = TRUE, cx = "left", xlab = "Latent Trait", cex.main = 1.5, cex.lab = 1.3, cex.axis = 1.1)
    - = each item, one by one, check response patterns

SS **Test** & **Item** **I**nformation **C**urves (English S1)

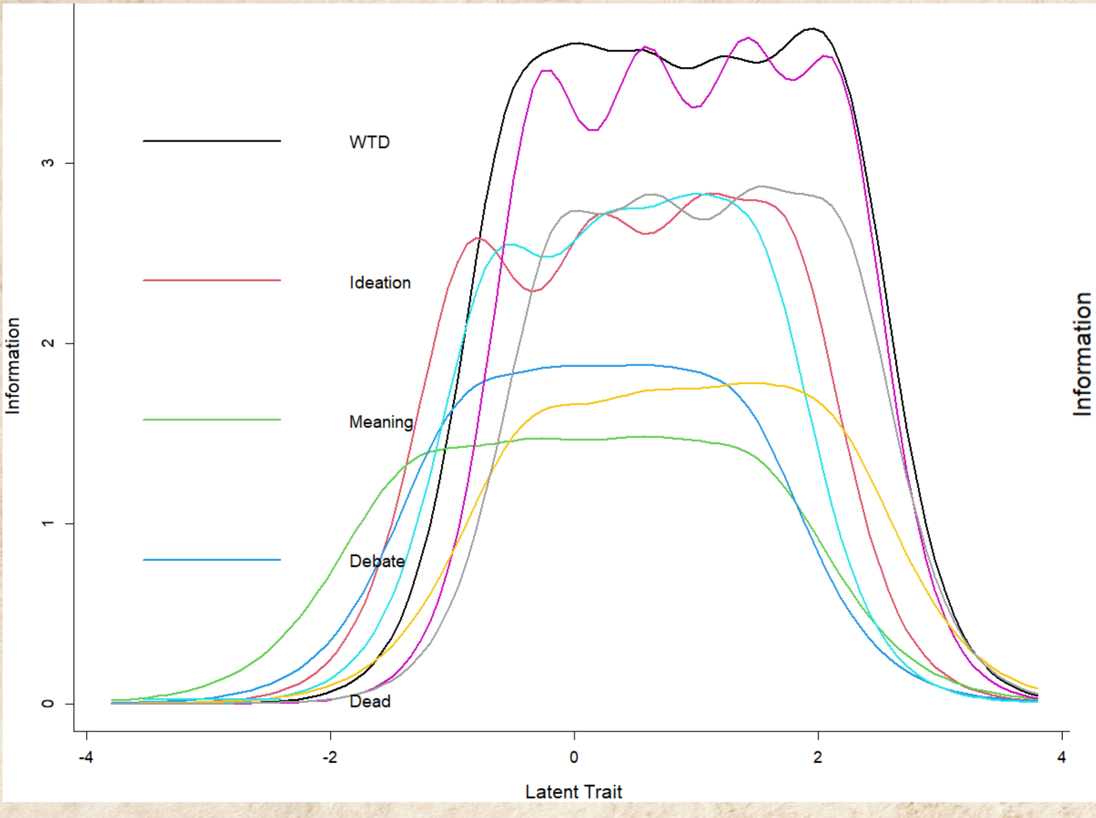


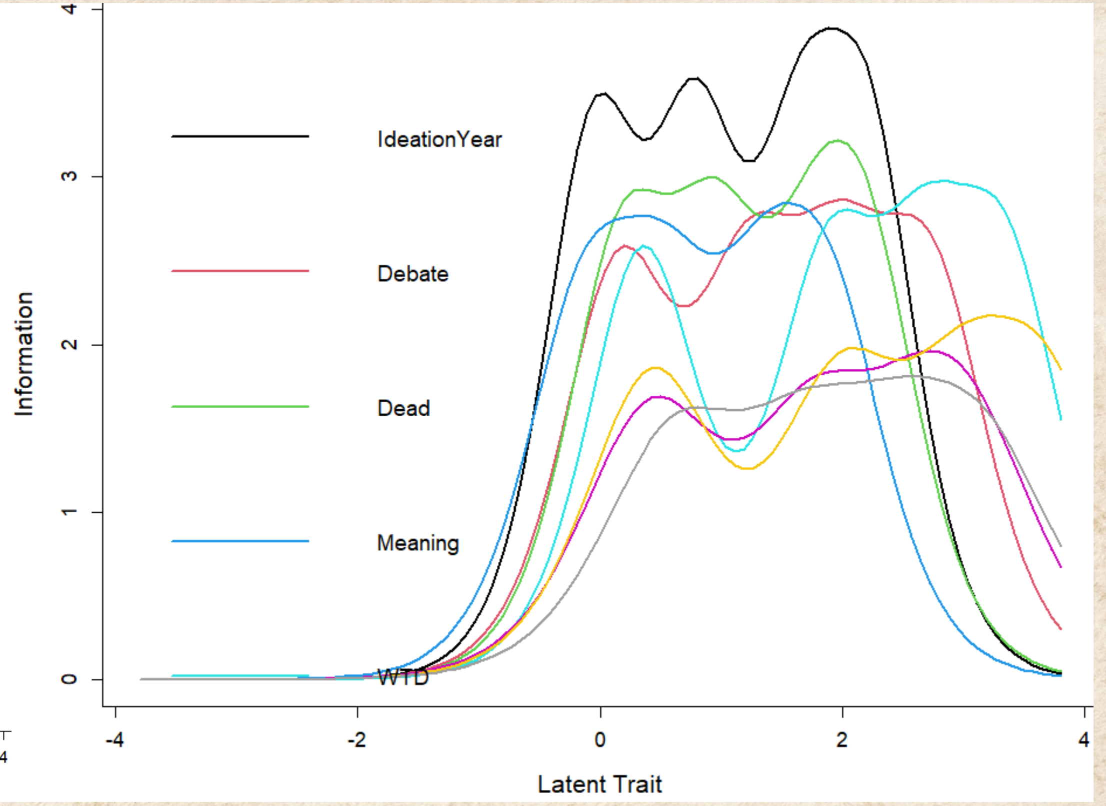


The **Test IC**, left, shows the total information captured on theta, for the full test (Suicidality Scale). Most information, under the line, comes from low-moderate to moderate-high levels of theta, with less information and more error at tails.

The **Item IC**s, right, show information captured on theta by each item. Some items show strength at low or high tails, no two are the same. Even if two items overlapped, they may provide unique information at theta levels.

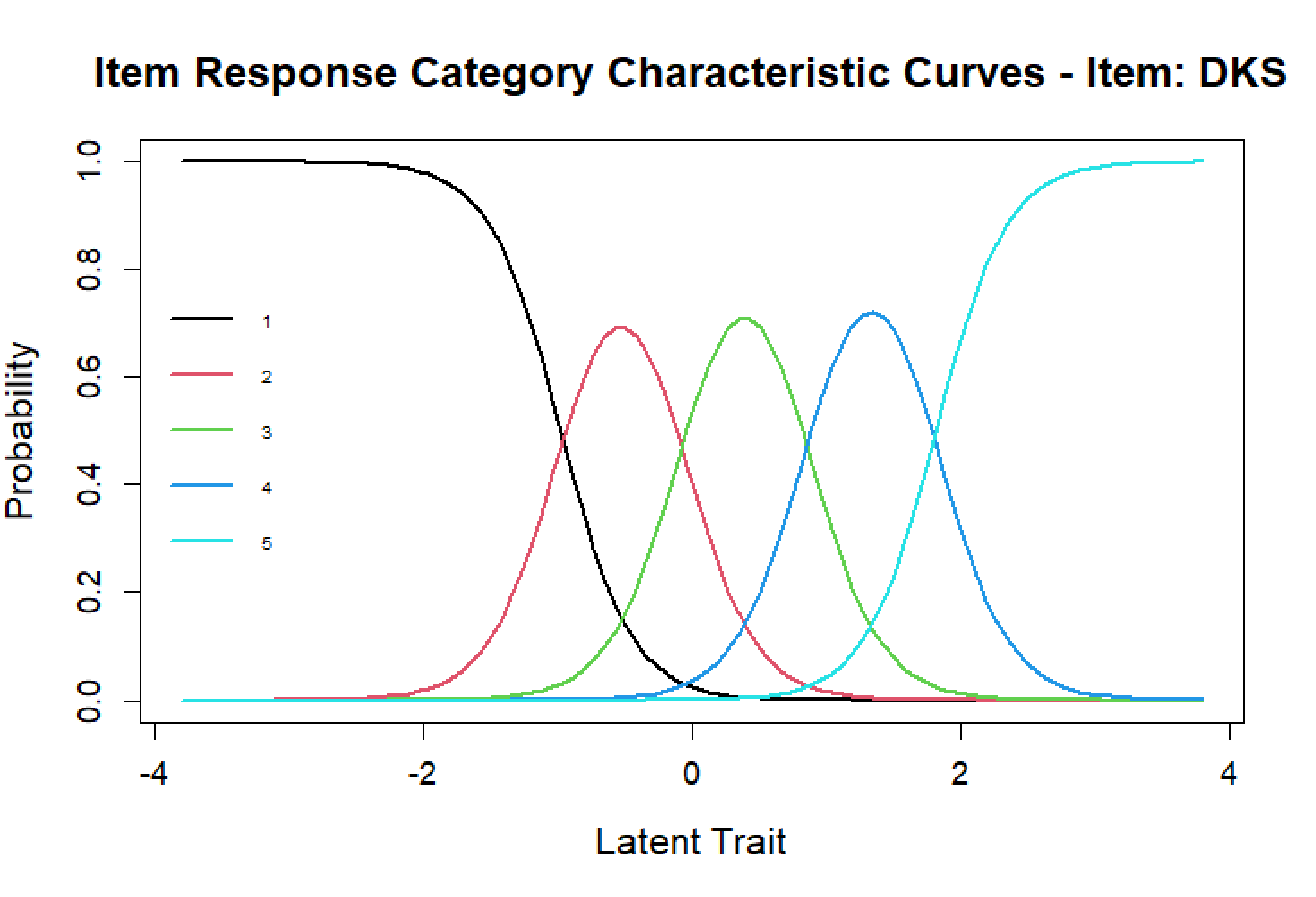
Item **C**haracteristic **C**urves (English S2, S3)





These **ICC**s are from smaller samples but still provide good information on how items capture unique information on theta. These graphs can also be used to help identify weak items in item pools.

Item **R**esponse **C**ategory **C**haracteristic **C**urve – Desire to Kill Self



This graphic illustrates ***b* coefficients** (item response cutpoints), indicating the *theta location* of response boundaries. E.g., the intersection between response 2 and response 3 (where the red and green lines intersect) $b_2 = -0.09$. We see the information, area under each line, each response captures.

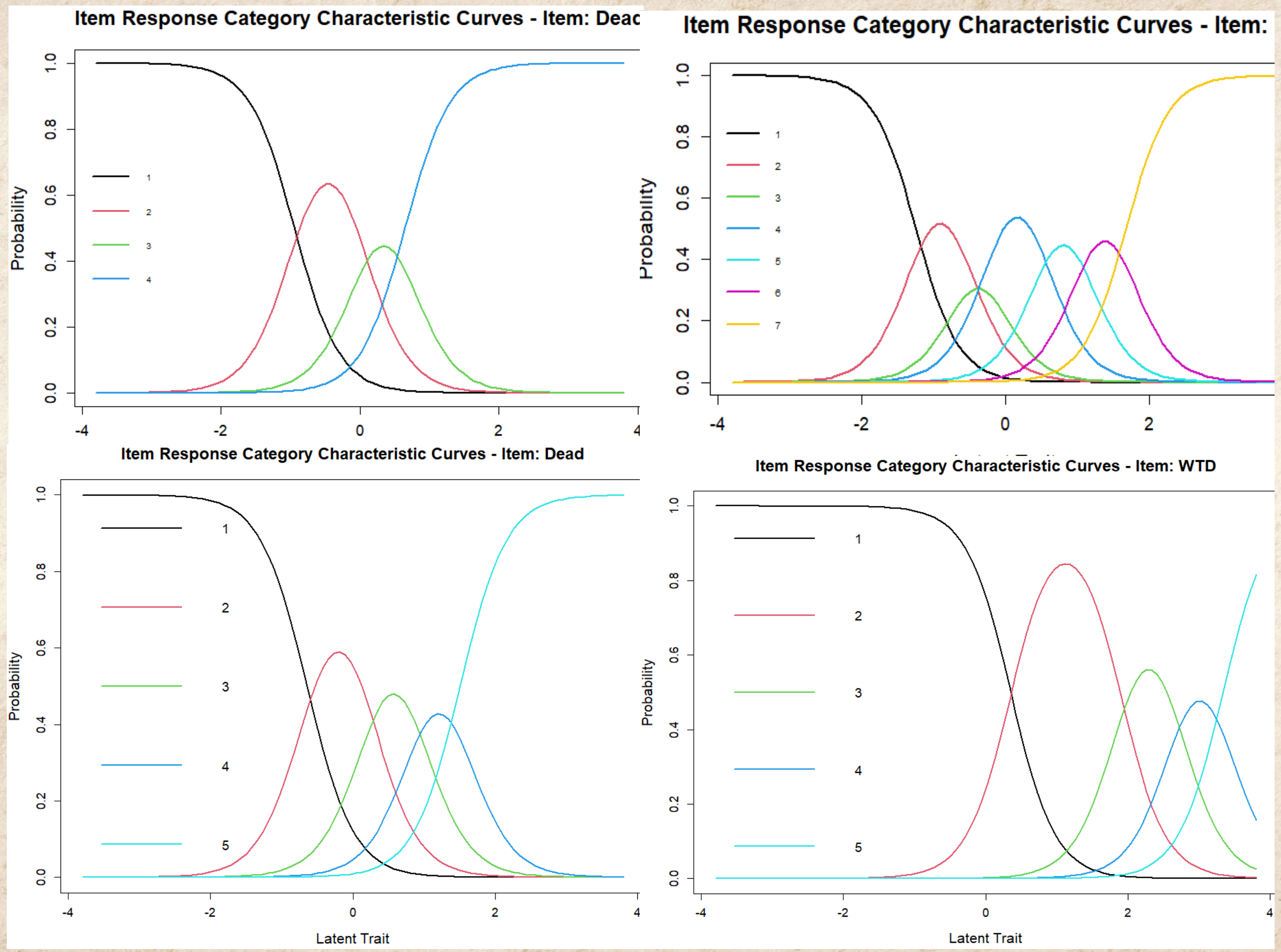
Compare b coefficients across items. They do not necessarily match, indicating item responses are not equivalent.

Can test **monotonicity** – responses consistently increase/decrease along theta. This item shows a near-textbook-level perfection. Nice orderly curves, lines nearly the same height and width.

Item **R**esponse **C**ategory **C**haracteristic **C**urves (English S1: S2/S3)

S1: **4-point** response (PHQ-9)

S2: **5-point** response, **reworded** Improvements in several stats



S1: **7-point** response
Note the 3rd response is under-endorsed; 7 points may be too many

S3: **5-point** response
Improvement in **monotonicity,** response-validity, but lacks high-theta data

Dead **WTD**

Putting it together – SS **Item** Stats

Item	Graded response model				Clus	FA		BFA	
	b_l	b_u	a	IF	L	L	h^2	g	h^2
DKS	-0.88	1.84	3.79	13.19	.92	.93	.86	.93	.89
WTD	-1.20	1.61	3.29	12.33	.91	.91	.83	.92	.85
RFD	-0.89	1.98	2.64	8.15	.83	.83	.68	.85	.73
Ideation-year	-1.55	0.44	2.81	7.46	.84	.84	.70	.79	.83
Dead	-0.96	0.61	3.01	6.93	.85	.85	.72	.83	.75
Predict	-0.99	1.33	2.40	6.39	.83	.83	.69	.83	.72
Debate	-1.26	0.60	2.57	6.34	.85	.85	.72	.81	.80
Meaning	-1.19	0.66	2.36	5.33	.82	.82	.67	.83	.73

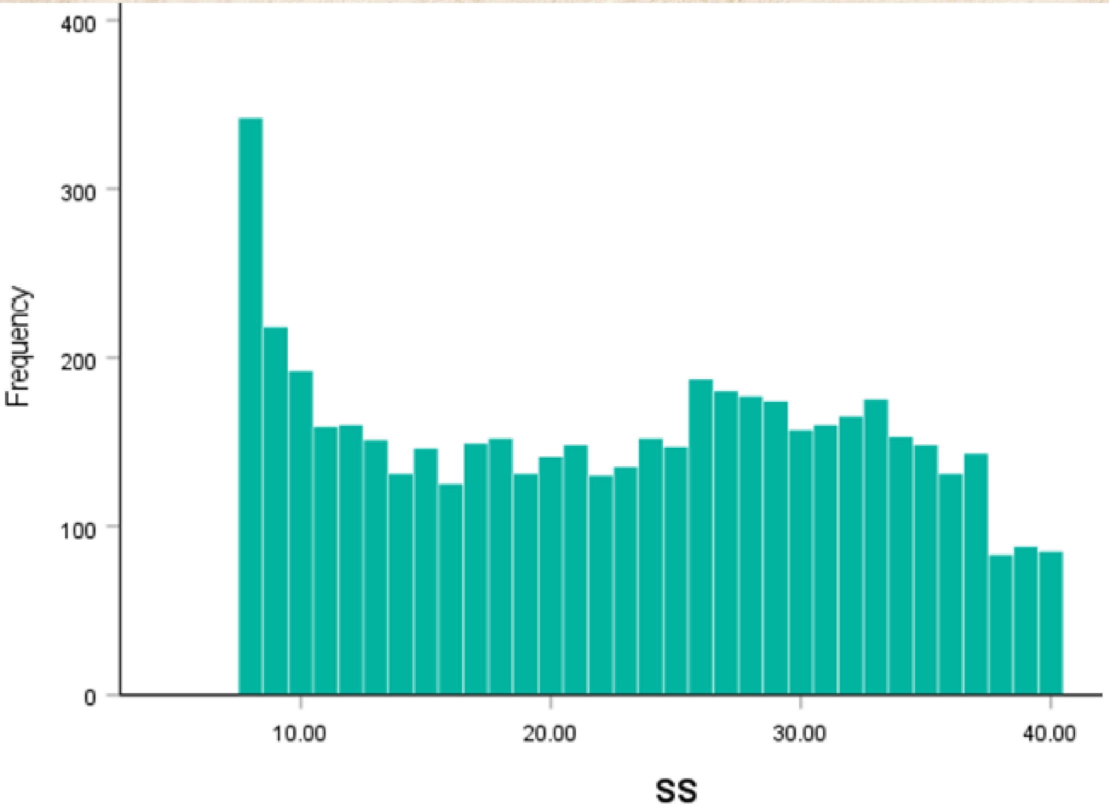
Here are detailed statistics of final SS items. All items showed strengths across all analyses.

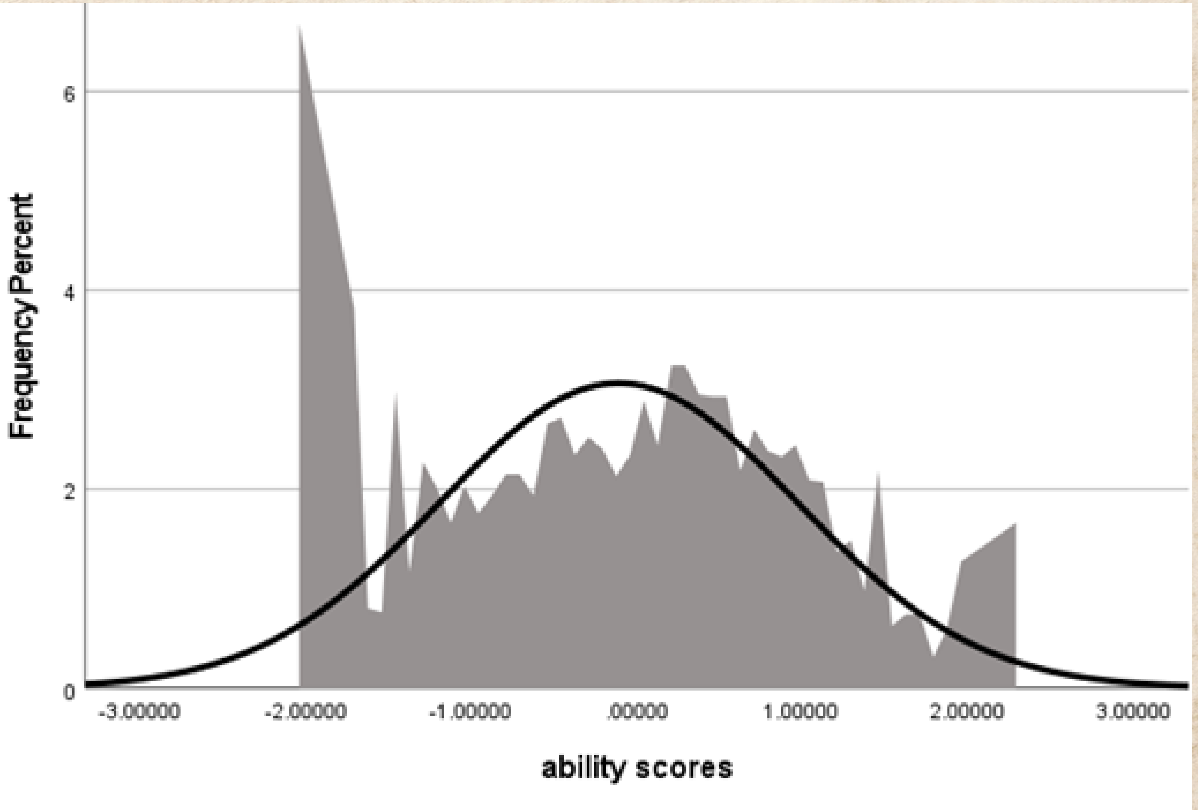
In green, we see two items that appear equal through some analyses. However, considering GRM and BFA, they have different strengths.

IRT – GRM – **Ability Scores** - Demonstration

- ltm package – **Factor/ability scores** – PROMIS© ‘response pattern scoring’
- (ensure **data is sorted by ID first!)**
- “options(**max.print** = 99999)”
  - > “fit2 <- grm(scaleA, , IRT.param = TRUE, constrained = FALSE)”
  - > “fs <- factor.scores(fit2, resp.patterns = data)”
  - > “sink('data.csv’)” data = your name for the new file, as you like
  - > “fs”
  - > “sink()”
- Locate ‘data.csv’ file, copy ability scores (scaleAz1) and se (scaleAse.z1) paste into dataset

SS **Sum** & **Ability** Scores (English S1)

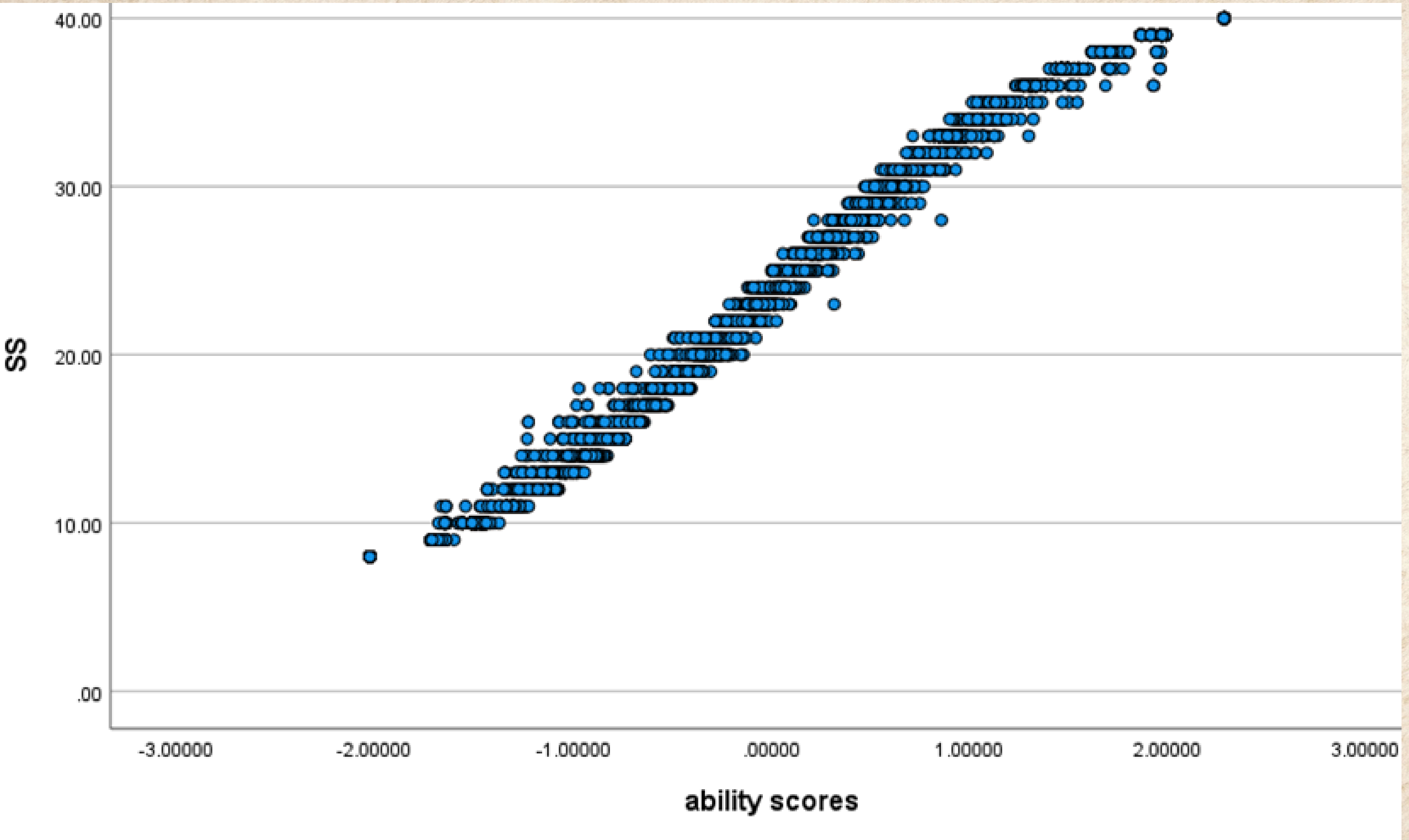




This histogram of SS **sum scores** shows good coverage of the trait.

This graph of SS **ability scores**, from the same dataset, shows many more score points and a more normal distribution.

Scatterplot SS Sum Scores & Ability Scores (English S1)



Y-axis = **sum** scores
X-axis = **ability** scores

Note that a sum score of 20 shows a range of ability scores, as do other sum scores. Highlighting better precision of ability scores through response pattern scoring.

IRT – GRM – **DIF** – Demonstration

- lordif package – **DIF**
  - Differential item functioning: do items assess the latent trait equivalently, or not (DIF), across groups (e.g., genders, age groups, ethnicity)
  - Currently, categories are limited to **2 or 3**. There need to be sufficient cases across levels of theta to determine DIF
- > “gender <- study1[, 36]”
  - Gender = new variable name (DIF variable); 36 = column in dataset for gender
- > “difgender <- lordif(scaleA, gender, criterion = "R2")”
  - Difgender = new variable name (DIF test of gender variable)

DIF Output

```
lordif(resp.data = data, group = diagnosis, criterion = "R2")
  Number of DIF groups: 2
  Number of items flagged for DIF: 0 of 8
  Items flagged:
  Number of iterations for purification: 1 of 10
  Detection criterion: R2
  Threshold: R-square change >= 0.02
DFIT Analysis
Group: 0
Iteration: 85, Log-Lik: -28885.141, Max-Change: 0.00010
 (mirt)
Group: 1
Iteration: 99, Log-Lik: -14918.401, Max-Change: 0.00008
 (mirt)
DTF (1) = 0.017
```

For DIF, we set the criterion for a meaningful difference in assessment at R2 < .02. These results show that self-reported diagnosis for a mental disorder status (yes/no) showed no evidence of DIF (0 of 8 items) or DTF.

Internal Consistency – **Omega – ω**

- coefficientalpha package
  - **Disconnect** psych package!

- Obtain **bootstrapped** robust omegas
  - Internal consistency (similar to Cronbach's alpha)
  - This will yield coefficient ω, and bootstrapped 95% CI intervals
  - You can just report the 95% CI, they are most important

- > "omega <-bootstrap(scaleA, type='omega', nboot=10, plot=TRUE)"

McDonald's ω Output

```
omega <-bootstrap(ss, type='omega', nboot=10, plot=TRUE)
```

The estimated omega is 0.9668936

Its bootstrap se is 0.002

Its bootstrap confidence interval is [0.964 , 0.968]

Here, we calculate the robust omega coefficient, and even more important – the bootstrapped 95% CI. You can just report the CI if you like.

Putting the Results Together

Basic **Interpretation** of Statistics

- Model fit and common trait variance = TLI, Fit, V (variance), ECV (explained common variance), ω_h – Closer to 1.0 is best
  - Each of these gives some indication of how well the scale fits the model
  - If the measure is unidimensional, these results should help confirm that
- Model error = RMSEA, RMSR (should be near 0, depending on metric)
- **Loadings** – closer to 1.0 is strong; < .60 may be concerning
- Communalities (h^2) – closer to 1.0 is good; < .50 may be concerning
- Internal **consistency** = ω, close to 1.0 is good; < .85 may be concerning
- These statistics may be best understood by comparing metrics across scales and studies. Please see References for expert guidance!

Scale **EFA** Results

We conducted psychometric analyses on all scales used in our studies, including bifactor analysis, GRM, HCA. Here, we present EFA and omega statistics.

These can be compared across studies, and with the Suicidality Scale.

More details in the SS English manuscript.

	Minimum residual factor analysis					ω
Study/scale	TLI	RMSEA	*V*	Loading	h^2	95% CI
Study 1 (*N* = 5115)						
SWLS	.98	.09	.65	.68 - .90	.46 - .81	[.88, .89]
PHQ-9	.89	.12	.56	.65 - .86	.42 - .75	[.89, .90]
PHQ-8	.91	.12	.55	.66 - .83	.44 - .69	[.87, .88]
DASS-Anxiety	.96	.09	.63	.63 - .88	.40 - .77	[.89, .90]
DASS-Depression	.94	.14	.74	.81 - .88	.65 - .78	[.93, .93]
C-SSRS-10	.82	.21	.70	.64 - .96	.41 - .92	[.87, .88]
C-SSRS-5	.98	.06	.54	.56 - .83	.31 - .69	[.85, .86]
SABCS-m	.87	.20	.65	.54 - .87	.29 - .76	[.91, .92]

SS – **Scale Statistics** (English)

These results show multimodel scale statistics. We see high fit, *mostly* low error, high variance explained, high internal consistency – across models and studies.

More details in the manuscript.

	Cluster		Factor Analysis			Bifactor Analysis			ω
Study	Fit	RMSR	TLI	RMSEA	V	ω_h	ECV	RMSEA	95% CI
S1	.98	.03	.94	.12	.74	.94	.92	.05	[.96, .96]
S2	.99	.03	.95	.14	.84	.93	.91	.06	[.96, .97]
S3	.99	.04	.89	.24	.87	.93	.87	.02	[.96, .97]

GRM Table

Graded Response Model Analyses of Suicidality Scale Items

Item	b_1	b_2	b_3	b_4	b_5	b_6	a	IF
DKS	-0.28	0.58	1.39	2.15	—	—	3.67	12.22
WTD	-0.52	0.03	0.60	1.23	1.84	2.22	3.52	9.41
RFD	-0.14	0.65	1.47	2.17	—	—	3.17	11.11
Ideation	-0.97	-0.18	0.63	1.27	—	—	3.60	11.97
Dead	-0.66	0.11	0.61	1.12	—	—	3.53	10.66
Predict	-0.24	0.51	1.19	1.89	—	—	2.42	6.43
Debate	-1.32	-0.43	0.36	1.08	—	—	2.79	7.81
Meaning	-1.53	-0.43	0.49	1.29	—	—	2.34	6.92

This is a more complete GRM table. We see the individual item response threshold levels for each item, the b coefficients. b1 shows where responses 1 and 2 cross. The b statistic relates to the level of theta, where those responses cross. Lower values indicate lower levels of the latent trait. We can see that some capture information on theta at lower levels than others (e.g., Meaning), some at higher levels (e.g., DKS). We also see that items can differ in many ways on thresholds.

a = discrimination, how well the item discriminates test-takers on theta

IF = information function, the amount of information captured on theta

Problems with Scores

- **Sum** scores – not precise, not individualized, excess error
- **Cutoffs**/cut points – rely on all items being equal
  - Not valid, excess error
- **Ability** scores – complicated, time consuming, often unavailable

- How can we **improve** latent trait scores – for usability & interpretation**?**

T-Scores

- T-scores are popular and easy to interpret (M = 50, SD = 10)
  - A T-score of 70 on a suicide risk assessment = two standard deviations worse than the average person assessed
- Can create T-scores from ability scores by including **ranges**
- Steps
  - Calculate z-scores (sum scores and ability scores)
  - T = 50 + (10 * z-score)
  - T-scores: SD = 10.0; M = 50.0
  - SE = SD/√n
  - N = sample size (should be large, diverse, representative for *official* T-scores)
  - 95% CI = 1.96 * SE
- Colombia study: SE = 10/ √313 = 0.57
  - (1.96 * SE) = 1.12
  - Best estimate = T-score +/- 1.12
- **Interpretations** – T-score = 49 – 51, about *average* suicidality (of community)
  - T-score = 59 – 61, one standard deviation above *average*, significantly higher symptoms
  - T-score = 79-81, three SDs above *average*, extreme symptoms

Approximate Ability T-Scores

- Not possible to match single sum score with a specific ability score
- When comparing with sum scores, these will have a **range**
- In addition, there will be **95% CI**, increasing the range
- We can present tables with sum scores = approximate ability T-score ranges
  - See PROMIS manuals for additional examples

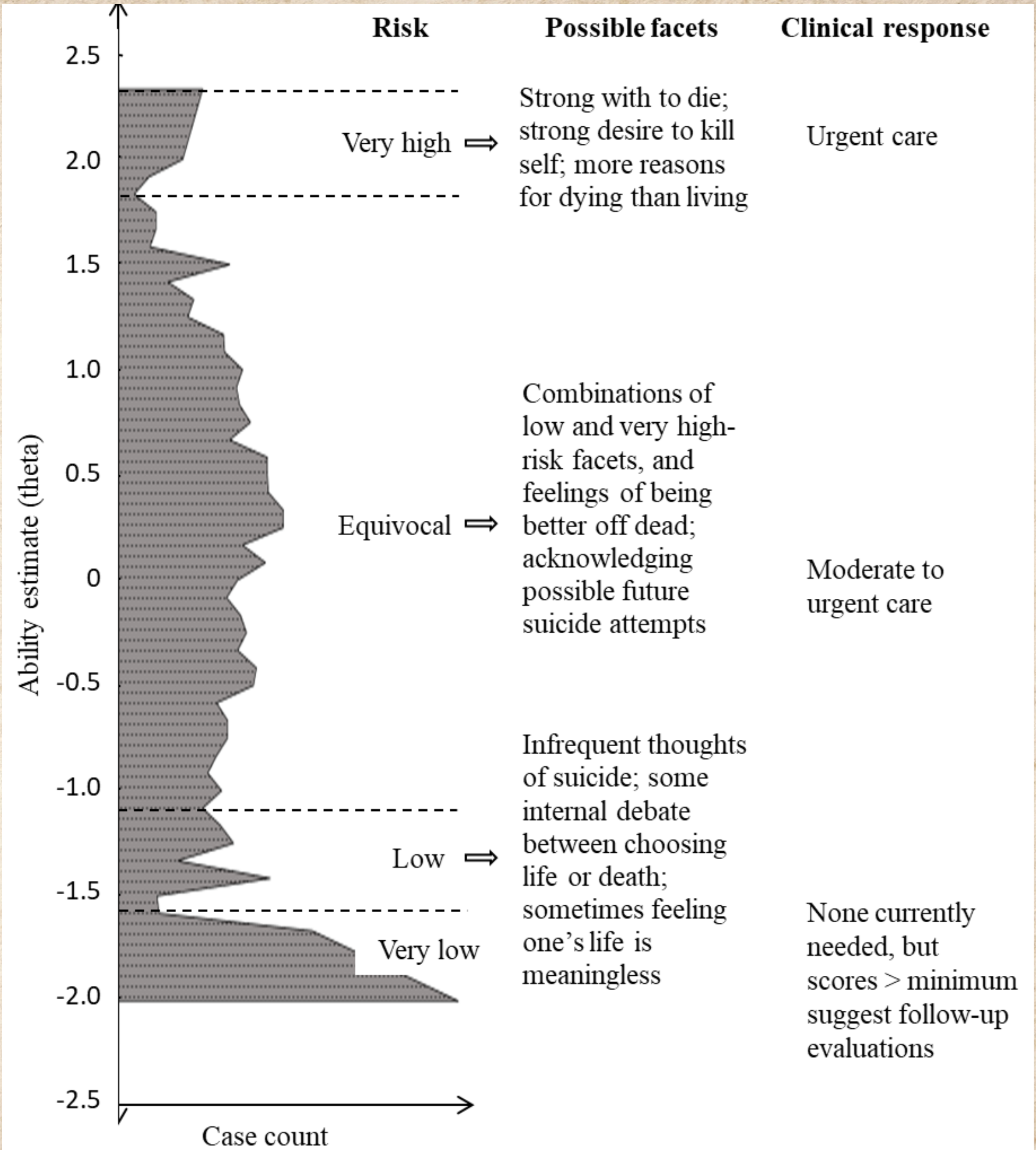
- **Example**: sum score = 40 ≈ *ability* T-scores 68 – 71
  - +/- 1.12, Approximate Ability T-score = 67 – 72
  - Sum scores will show **overlapping** ability T-scores
  - These are **approximate** ability scores. A *true* ability score might fall out of this range

Clinical Guide

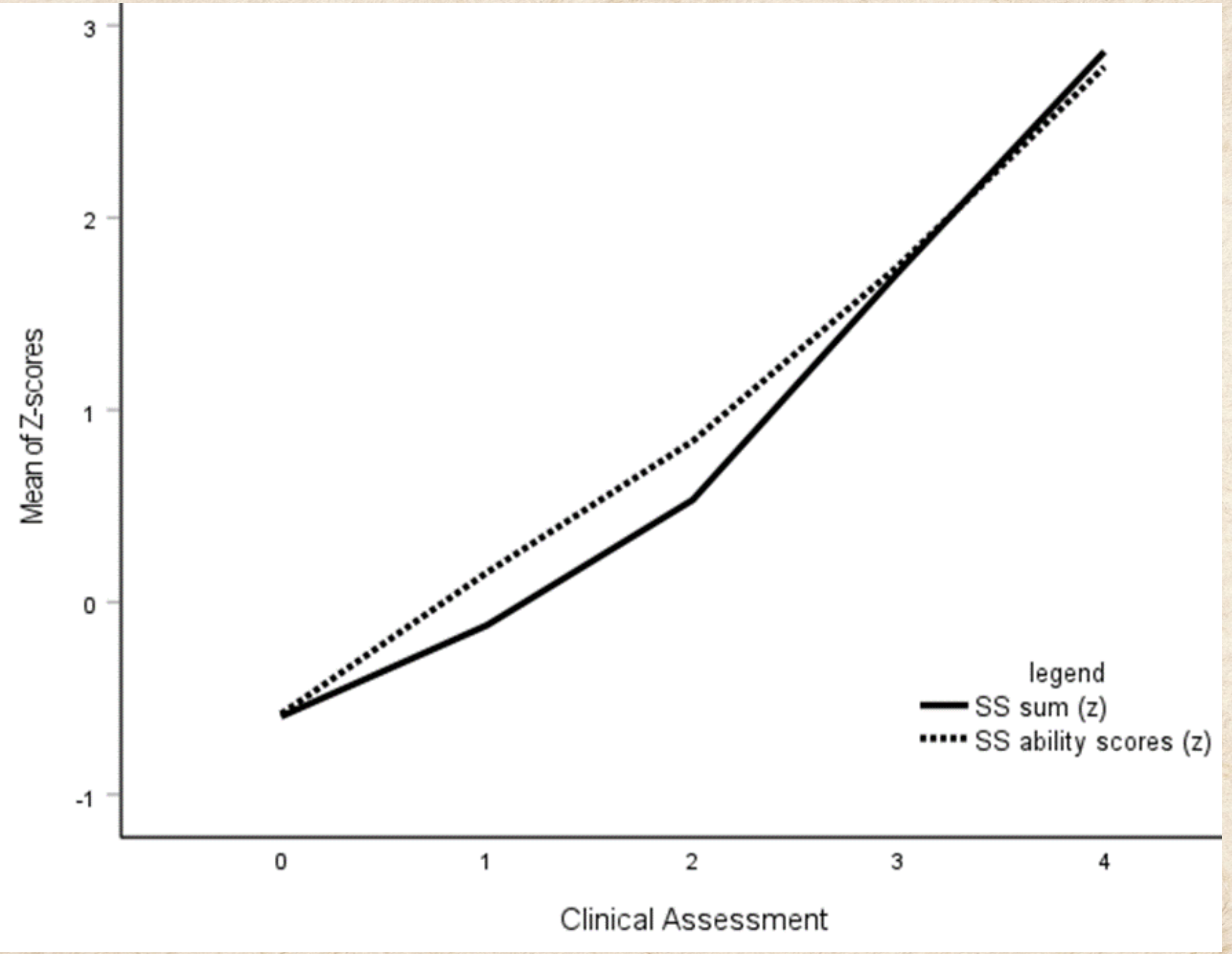
Sum Scores	Sum T-Scores	GRM T-Scores
48	80	79
47	79	76-77
45	77	73-74
40	72	68-71
35	68	65
29	62	61-63
21	55	56-59
15	**50**	51-56
11	46	47-51
10	45	47-50
9	44	46-48
8	43	41
95% CI =	+/- 1.12	+/- 1.12

The left table shows Colombian sample sum scores can translate to T-scores, and how these compare to GRM T-scores.

On the right, we have English S1 ability (GRM) scores related to suicidal facets and clinical directions.



Clinician Ratings – Colombian sample



This plot shows Clinical decisions, x-axis, ordered from 0 = no treatment to 4 = immediate hospitalization
The lines represent z-scores of SS sum scores and SS ability scores.

Note that the clinical decisions are more closely correlated with the ability scores at 1 and 2. This may indicate that clinical decisions are based on information beyond the sum scores, making them more valid, and closer to the more precise ability scores.

Applying Knowledge

Choosing Latent-trait Measures

- **Consider**
  - IRT-tested > concise, highly informative
  - DIF-tested > works equivalently in different demographic groups
    - E.g.: PROMIS https://www.healthmeasures.net/explore-measurement-systems/promis
    - PROMIS has several validated scales, **free** to use but **not modifiable**
  - **CAT** (computer adaptive testing), important direction, **but** beware of **GIGO**
    - CAT requires highly valid instruments to be effective
- **Requirements**
  - Assesses **individual** traits, not **group** factors
    - E.g., SADPERSONS includes items on sex, age – which may indicate group risk, *not* personal
  - Scale should include critically tested items, evidence-based
  - Does not emphasize cutoff scores (low, medium, high-risk)
    - E.g., DASS-21, emphasizes ranges, subjective interpretation of sum scores
  - Does not rely on: Cronbach's alpha > .70; Factor loadings > .32+; CFA fit/error
    - Provides more item/scale statistics and emphasizes strengths/weaknesses

Proactive Directions

- Support/Promote **Open Science**
  - e.g., set up accounts, contribute 1 thing
- Support/Promote **SDG**s: How can your work contribute?
  - e.g., women & minority leadership; global connections; sharing knowledge
- Support/Promote **Free & Valid Instruments**
  - Choose scales/measures/instruments carefully
  - Mention/cite **free culture** licensed resources
- **All Science** ~~is~~ *needs to be* **Local**
  - Support/Promote **localization** of skills, instruments, etc.

Contact – Keith Harris keithharris@csu.edu.au kmh.psyc@gmail.com

Original talk:
https://charlessturt.zoom.us/rec/share/Vl78EZPEvEheOTul_YCmgBNDsaIb24oxgLCyuHrTQWWUMoiKxKx2hp1JfqHK_ggi.eWwiLPW6odsn8LcR?startTime=1658440991000

Charles Sturt University

Extra **Resources**

Do **1** thing**!**

Open science sites

Research ethics/guidelines

References

Get started: Do 1 Thing!

- **Contact** someone through OSF or other
- **Preregister**, post preprint of, your research @ OSF or similar
- **Post** some good work you have done, but perhaps won't be a published paper
- An **example** of **1** thing
  - A revised poster presentation, published through figshare, including a doi, and publicly available
  - https://figshare.com/account/projects/4566/articles/20237100 Reasons for Dying (RFD)

Open Science-related Sites

License your paper/data/poster/etc. as **Free Cultural**

- We copyrighted the Suicidality Scale as **CC BY 4.0** (free cultural)

https://creativecommons.org/

For open data, preprints, preregistration, etc.: https://osf.io/

Similar to OSF: https://figshare.com/

More **Open** Science Links

- Start by getting an ORCID, setting up accounts at OSF, Scholar, etc.
- Researcher ID: https://**orcid**.org/
- Google Scholar: https://scholar.google.com
- ResearchGate: https://www.researchgate.net/
  - Example project: https://www.researchgate.net/project/Assessing-Suicidality-Development-of-State-of-the-Science-Measures-of-Suicide-Risk-Across-Languages-and-Cultures
- More free & open source **statistics** packages: Jamovi: https://www.jamovi.org/
  - JASP: https://jasp-stats.org/download/
- There are **many more** open science sites!

Research Ethics/**Guides**

- The Helsinki Declaration: **Ethics in medical research** (see World Medical Association)
  - https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
- **Data management**: FAIR Wilkinson et al. (2016)
  - https://www.nature.com/articles/sdata201618#citeas
- **Author** roles: https://credit.niso.org/
- **TOP** open research:
  - Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science, 348(6242), 1422-1425. doi:10.1126/science.aab2374*
- Nature's *Scientific Data* offers guidance, with links, on **data repositories**:
  - https://www.nature.com/sdata/policies/repositories#general
  - Nature's **data policy** page: https://www.nature.com/sdata/policies/data-policies
- PLOS also offers **data** info and links: https://journals.plos.org/plosone/s/recommended-repositories

Select **References** for Surveys, Psychometrics, etc.

- Bentley, J. P., & Thacker, P. G. (2004). The influence of risk and monetary payment on the research participation decision making process. *Journal of Medical Ethics*, *30*(3), 293-298. https://doi.org/10.1136/jme.2002.001594
- Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, *34*(3), 277-313. https://doi.org/10.1207/S15327906MBR3403_1
- Bland, J. M., & Altman, D. G. (2015). Statistics notes: Bootstrap resampling methods. *BMJ (Online)*, *350*, h2622. https://doi.org/10.1136/bmj.h2622
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425-440. https://doi.org/10.1007/s11336-006-1447-6
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391-418. https://doi.org/10.1177/0013164404266386
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
- Décieux, J. P., Mergener, A., Neufang, K. M., & Sischka, P. (2015). Implementation of the forced answering option within online surveys: Do higher item response rates come at the expense of participation and answer quality? *Psihologija*, *48*(4), 311-326. https://doi.org/10.2298/PSI1504311D
- DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed., Vol. 26). Sage.
- Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 1-17. https://doi.org/10.1186/2193-1801-2-222

Select References

- Dupuis, M., Meier, E., & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods*, *51*(5), 2228-2237. https://doi.org/10.3758/s13428-018-1103-y
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Assoc.
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456-465. https://doi.org/10.1177/2515245920952393
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. https://doi.org/10.1177/1948550617693063
- Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(1), 72. https://doi.org/10.1186/s12916-015-0325-4
- Gnambs, T., & Kaspar, K. (2014). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*(4), 1237-1259. https://doi.org/10.3758/s13428-014-0533-4
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930-944. https://doi.org/10.1177/0013164406288165
- Hanel, P. H. P., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PloS One*, *11*(12), e0168354. https://doi.org/10.1371/journal.pone.0168354
- Harris, K. M., Lello, O. D., & Willcox, C. H. (2017a). Reevaluating suicidal behaviors: Comparing assessment methods to improve risk evaluations. *Journal of Psychopathology and Behavioral Assessment*, *39*(1), 128-139. https://doi.org/10.1007/s10862-016-9566-6
- Harris, K. M., Syu, J. J., Lello, O. D., Chew, Y. L. E., Willcox, C. H., & Ho, R. H. M. (2015). The ABC's of suicide risk assessment: Applying a tripartite approach to individual evaluations. *PloS One*, *10*(6: e0127442), e0127442. https://doi.org/10.1371/journal.pone.0127442
- Hofmans, J., Theuns, P., & Mairesse, O. (2007). Impact of the number of response categories on linearity and sensitivity of self-anchoring scales: A functional measurement approach. *Methodology*, *3*(4), 160-169. https://doi.org/10.1027/1614-2241.3.4.160
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828-845. https://doi.org/10.1037/a0038510

Select References

- Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 166-184. https://doi.org/10.1177/2515245919882903
- Joinson, A. N., Woodley, A., & Reips, U.-D. (2007). Personalization, authentication and self-disclosure in self-administered Internet surveys. *Computers in Human Behavior*, *23*(1), 275-285.
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847-865. https://doi.org/10.1093/poq/nfn063
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73-79. https://doi.org/10.1027/1614-2241.4.2.73
- MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36*(4), 611-637. http://www.scopus.com/inward/record.url?eid=2-s2.0-0035648558&partnerID=40&md5=fd0b9f8efa5bb0b38a9b63918c75c747
- MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, *90*(4), 710-730. https://doi.org/10.1037/0021-9010.90.4.710
- Madsen, J., & Harris, K. M. (2021). Negative self-appraisal: Personal reasons for dying as indicators of suicidality. *PloS One*, *16*(2), e0246341. https://doi.org/10.1371/journal.pone.0246341
- Mair, P. (2018). *Modern psychometrics with R*. Springer. https://doi.org/https://doi.org/10.1007/978-3-319-93177-7
- Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research*, *51*(5), 698-717. https://doi.org/10.1080/00273171.2016.1215898
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287-2305. https://doi.org/10.3758/s13428-020-01398-0
- Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388. https://doi.org/10.1177/1094428104268027
- Meijer, R. R., & Egberink, I. J. L. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement*, *72*(4), 589-607. https://doi.org/10.1177/0013164411429344

Select References

- Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, *17*(4), 351-363.
- Nering, M. L., & Ostini, R. (2011). *Handbook of polytomous item response theory models*. Routledge.
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1-11. https://doi.org/10.1016/j.jrp.2016.04.010
- Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, *14*(3), 548-570. https://doi.org/10.1177/1094428110368562
- Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*(2), 208-225. https://doi.org/10.1037/met0000126
- Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS®): Depression, anxiety, and anger. *Assessment*, *18*(3), 263-283. https://doi.org/10.1177/1073191111411667
- Powell, M. A., & Smith, A. B. (2009). Children's participation rights in research. *Childhood*, *16*(1), 124-142. https://doi.org/10.1177/0907568208101694
- PROMIS. (2016). *PROMIS item bank v2.0, Emotional Support Short Form 6a*. PROMIS Health Organization and PROMIS Cooperative Group. https://www.healthmeasures.net/index.php?Itemid=992
- R Core Team. (2021). *R: A language and environment for statistical computing*. In (Version 4.1.1) R Foundation for Statistical Computing. https://www.R-project.org/
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173-184. http://www.scopus.com/inward/record.url?eid=2-s2.0-0031486705&partnerID=40&md5=8bb526dfa831ee912767d2b65545189b
- Raykov, T. (1998). On the use of confirmatory factor analysis in personality research. *Personality and Individual Differences*, *24*(2), 291-293. https://doi.org/10.1016/S0191-8869(97)00159-1
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154. https://doi.org/10.1007/s11336-008-9102-z

Select References

- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223-237. https://doi.org/10.1080/00223891.2015.1089249
- Samejima, F. (2004). Graded response model. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 77-82). Academic Press.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120. https://doi.org/10.1007/s11336-008-9101-0
- Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*. https://doi.org/10.1007/s11336-021-09789-8
- Stieger, S., Reips, U. D., & Voracek, M. (2007). Forced-response in online surveys: Bias from reactance and an increase in sex-specific dropout. *Journal of the American Society for Information Science and Technology*, *58*(11), 1653-1660. https://doi.org/10.1002/asi.20651
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859-883. https://doi.org/10.1037/0033-2909.133.5.859
- Vicente, P., & Reis, E. (2010). Using questionnaire design to fight nonresponse bias in web surveys. *Social Science Computer Review*, *28*(2), 251-267. https://doi.org/10.1177/0894439309340751
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3(1), 160018. doi:10.1038/sdata.2016.18*
- World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, *310*(20), 2191-2194.
- Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409-428. https://doi.org/10.3758/s13428-018-1055-2
- Yentes, R. D., & Wilhelm, F. (2018). *Careless: Procedures for computing indices of careless responding*. In (Version 1.1.3) [R].
- Zhang, Z., & Yuan, K.-H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement*, *76*(3), 387-411. https://doi.org/10.1177/0013164415594658
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133. https://doi.org/10.1007/s11336-003-0974-7