Title

Validity and Validation: A Pragmatic Path Forward

Author

Wolf, Melissa Gordon

Publication Date

2022

UNIVERSITY OF CALIFORNIA

Santa Barbara

Validity and Validation: A Pragmatic Path Forward

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Education

by

Melissa Gordon Wolf

Committee in charge:
Professor Andrew Maul, Chair
Professor Karen Nylund-Gibson
Professor Ann Taves
Professor Daniel McNeish
Professor Nancy Collins

December 2022

The dissertation of Melissa Gordon Wolf is approved.

Karen Nylund-Gibson

Ann Taves

Daniel McNeish

Nancy Collins

Andrew Maul

December 2022

Validity and Validation: A Pragmatic Path Forward

VITA OF MELISSA GORDON WOLF
Dec 2022

EDUCATION

Bachelor of Arts in Communications, University of Delaware, May 2009
Graduate Certificate in Measurement, Statistics and Evaluation, University of Maryland, College Park, May 2012
Master of Arts in Research Methods and Statistics, University of Denver, June 2017
Master of Arts in Education, University of California, Santa Barbara, March 2020
Doctor of Philosophy in Education, University of California, Santa Barbara, December 2022 (expected)

PROFESSIONAL EMPLOYMENT

2011-2012: Teaching Assistant, University of Maryland, College Park
2012-2015: Research Analyst, International Baccalaureate
2015-2016: Graduate Student Researcher, University of Denver
2016-2019: Graduate Student Researcher, University of California, Santa Barbara
2019-2021: Teaching Assistant, University of California, Santa Barbara
2021-2022: Data Consultant, New Tech Network
2022-Present: Programming Consultant, Arizona State University
2022-Present: User Experience Researcher, Microsoft

PUBLICATIONS

Wolf, M. G. & McNeish, D. (2022). dynamic: An R Package for Deriving Dynamic Fit Index Cutoffs for Factor Analysis. *Multivariate Behavioral Research*.

McNeish, D. & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*.

Boness, C.L., Helle, A.C., Miller, M.B, Wolf, M.G., & Sher, K.J. (2022). Who opts in to alcohol feedback and how does that impact behavior? A pilot trial. *Journal of Studies on Alcohol and Drugs, 83*(5), 640-645.

Wolf, M. G., Ihm, E., Maul, A., & Taves, A. (2022). Survey item validation. In S. Engler & M. Stausberg (Eds.), *Handbook of Research Methods in the Study of Religion (2nd ed.)*. Routledge.

McNeish, D. & Wolf, M. G. (2021). Dynamic Fit Index Cutoffs for Confirmatory Factor Analysis Models. *Psychological Methods*.

Clairmont, A., Wolf, M. G., & Maul, A. (2021). The prevention and detection of deception in self-report survey data. In U. Luhanga & G. Harbaugh (Eds.), *Basic Elements of Survey Research in Education: Addressing the Problems Your Advisor Never Told You About.* Charlotte, NC: Information Age Publishing.

McNeish, D., & Wolf, M.G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52, 2287–2305.

Luo, Y. & Wolf, M. G. (2019). Item parameter recovery for the two parameter testlet model with different estimation methods. *Psychological Test and Assessment Modeling, 61*(1), 65-89.

Ghafoori, B., Wolf, M. G., Nylund-Gibson, K., & Felix, E. D. (2019). A naturalistic study exploring mental health outcomes following trauma-focused treatment among diverse survivors of crime and violence. *Journal of Affective Disorders*, 245, 617–625.

Raines, T.C., Gordon, M., Harrell-Williams, L.M., Diliberto, R.A, & Parke, E.M. (2017). Adaptive skills and academic achievement in Latino students. *Journal of Applied School Psychology*, 245 - 260.

Gordon, M., VanderKamp, E. & Halic, O. (2015). Research brief: International Baccalaureate programmes in Title I schools in the United States: Accessibility, participation and university enrollment. https://www.ibo.org/globalassets/publications/ib-research/title-1-schools-research.pdf

Bergeron, L. & Gordon, M. (2015). Establishing a STEM pipeline: Trends in male and female enrollment and performance in higher level STEM courses. *International Journal of Science and Mathematics Education*, 1 - 18.

Gordon, M., & Bergeron, L. (2014). The use of multilevel modeling and the level two residual file to explore the relationship between Middle Years Programme student performance and Diploma Programme student performance. *Social Science Research*, 50, 147-163.

SOFTWARE

Wolf, M. G. & McNeish, D. (2020). Dynamic Model Fit (version 1.1.0.). [Software]. Available from www.dynamicfit.app

Wolf, M. G. & McNeish, D. (2020). dynamic: Model fit cutoffs. R package version 1.1.0. https://cran.r-project.org/web/packages/dynamic/index.html

AWARDS

Block Grant Dissertation Award, University of California, Santa Barbara, 2020
Department of Education Excellence Award for Research, University of California, Santa Barbara, 2019
Grad Slam Finalist, University of California, Santa Barbara, 2019
Block Grant Fellowship Award, University of California, Santa Barbara, 2018
Education Travel Grant, University of California, Santa Barbara, 2018 – 2019
New Tech Network Research Grant, Napa, CA, 2016
Block Grant Fellowship Award, University of California, Santa Barbara, 2016
University of Denver Graduate Student Travel Grant, 2016
University of Denver Scholarship Award, 2015
Dean’s Fellowship, University of Maryland, College Park, 2010 – 2011

FIELDS OF STUDY

Major Field: Quantitative Methods in the Social Sciences

Studies in Psychometrics with Andrew Maul

Studies in Statistical Modeling with Karen Nylund-Gibson and Daniel McNeish

ABSTRACT

Validity and Validation: A Pragmatic Path Forward

by

Melissa Gordon Wolf

In the social sciences, data is often generated from responses to self-report survey instruments. Thus, researchers are often concerned with investigating survey instrument quality, more commonly known as validity. Evidence of instrument quality is typically presented quantitatively using correlational methods such as confirmatory factor analysis (CFA), regression, and coefficient alpha. Rarely do researchers engage in extensive pretesting of survey items or present evidence that survey items are understood as intended by the population of interest. Despite potential ambiguity in score meaning, these assessments are commonly used in high stakes settings such as measuring treatment efficacy in medical trials or evaluating the suitability of applicants for careers.

In this three-paper dissertation, I highlight some flaws with the current approaches to scale validation and present two new methods that will hopefully lead to improvements in survey quality. I begin by introducing the Response Process Evaluation method; a standardized framework for iteratively pretesting multiple versions of survey items and generating individual item level validity reports. Next, I discuss the improper over-generalization of a set of approximate fit index cutoff values for CFA models, and introduce a simulation-based, model-specific alternative called Dynamic Fit Index (DFI) cutoffs which have been made easily accessible in a new point-and-click software. I conclude by reviewing several of the most used scales in education and psychology and examining the types of validity evidence presented in defense of their use. My goal is to encourage researchers to think more critically about validity evidence and consider revisiting best practices in validation.

Introduction

In the social sciences, validity can be thought of as the adequacy or appropriateness of the use of a scale for a particular purpose (Maul, 2018). Validity theory and the practice of validation has evolved considerably since the birth of psychological testing in the early 1900s, and modern day best practices can vary substantially both within and across fields (Chan, 2014; Slaney, 2017). Within the social sciences, validation is further fragmented by the *Standards for Educational and Psychological Testing*'s arguments based approach to validity, which recommends presenting up to five different types of validity evidence given the proposed interpretations and uses of an assessment (AERA, APA, NCME, 2014; Kane, 2013). Moreover, the "publish or perish" mantra of academia often necessitates publishing scales with insufficient validity evidence, resulting in an overabundance of scales and confusion about what qualifies as rigorous psychometric research.

The goal of this dissertation is to produce three papers that highlight some of these issues and introduce some practical courses of action that will hopefully improve the current state of affairs. This work builds upon existing calls to action (e.g., Borsboom, 2006; Fried & Nesse, 2015; Maul, 2017; McNeish et al., 2018; Michell, 2012) and introduces some ideas that will likely not solve the problems in psychological measurement but may offer researchers more confidence in interpretations made from psychological assessments. I hope to re-emphasize the importance of theory and move us away from an over-reliance on quantitative evidence. This is particularly significant given that the most commonly used statistical models (e.g., factor analysis and coefficient alpha) make implicit claims about the structure of the psychological attribute in question (i.e., that it is quantitative and can therefore be modeled as such). By re-centering theory, I hope to encourage researchers to present both qualitative and quantitative evidence of validity and (when appropriate) use the statistical model that best aligns with their organic beliefs of the structure of the property in question.

Background

The first attempt at quantifying and measuring human traits was made by Francis Galton (1822 – 1911) in the late 1800s. Around the turn of the 20th century, mental testing of intellectual abilities and aptitudes became popular, and with that came the need for methods to analyze the resulting data. Karl Pearson (1857 – 1936) and Charles Spearman (1863 – 1945) adopted Galton's correlational methods and helped establish correlational analysis as the primary approach to analyze this mental testing data. Thus, the earliest psychometricians simply assumed that psychological properties had an underlying quantitative structure that was measurable and set out to do so (Slaney, 2017). Alfred Binet (1857 – 1911), the creator of the first IQ test, slightly dissented from his peers in this perspective: he acknowledged that psychological properties might not be truly quantitative but was not bothered by it. Instead, he believed that the scores that people achieved on his IQ test could accurately rank order people in terms of intelligence, and thus could still be used as a measure of intelligence (Michell, 2012). Thus, the earliest psychometricians generally followed what Michell (2003) describes as the "quantitative imperative": the idea that quantity is necessary for measurement and measurement is necessary for the social sciences to be considered scientific.

As psychological testing became more widespread in the early 1900s it came under scrutiny, particularly because of noted inconsistency in test scores across instruments and testing occasions (Slaney, 2017). Spearman (1904a, 1904b), frustrated by the inconsistency in test scores and lack of progress in psychology, deduced that there must be some sort of error built into test scores that caused scoring irregularities. He determined that a person's true ability was equal to their observed test score plus some error and sought to parse out that error such that a more accurate estimate a person's ability could be computed. This became the foundation of classical test theory (CTT) and *reliability*, which can be presented as $\rho_{XX'} = \left[\frac{Var(T)}{Var(T)+Var(E)}\right]$, where $\rho_{XX'}$ is a correlation coefficient indicative of some form of repeated testing or parallel test, T is the true score, and E is the error term. The idea here was that test quality could be evaluated by the magnitude of the reliability coefficients and, in theory, true ability levels could be better estimated by parsing out the error from the observed score. As such, psychologists became accustomed to reporting reliability coefficients, such as the Pearson correlation coefficient, Spearman's ρ (if the data are rank ordered), or coefficient alpha. Reliability coefficients are commonly reported today, appearing in up to 94% of scale-development or scale-use articles (Cizek et al., 2008; Flake et al., 2017; Zumbo & Chan, 2014). Coefficient alpha is particularly popular, appearing in up to 93% of articles that report reliability coefficients (McNeish, 2018).

Around the 1920s, the concept of test *validity* became increasingly recognized as an important property of tests. Validity was originally defined as "the extent to which [tests] measure what they purpose to measure" (Buckingham, 1921, p. 274), and although this definition is widely considered to be outdated[1], it is still commonly found in contemporary literature (Slaney, 2017). Originally, psychologists were preoccupied by what is now known as evidence for validity based on the test content and an external criterion, which was often

[11] For a noteworthy dissent, see Borsboom et al., 2004.

presented through the use of content analysis to establish a logical link between the test content (e.g., academic achievement) and an external criterion (e.g., an IQ test). The popularity of the correlation coefficient in reliability research likely led to its use in validity research, where researchers began computing correlation coefficients between tests and related criteria. This “validity coefficient” became the dominant type of validity evidence since it was simple, concise, and quantitative in nature. However, the quality of the criterion inevitably came into question, especially when tests were found to correlate at similar magnitudes with many other external criteria (some of which were also psychological tests prone to measurement error), resulting in further reform of validity theory (Slaney, 2017).

With the advent of the common factor model in the 1930s (Thurstone, 1931), psychologists became interested in the structure of the property of interest, i.e., the extent to which test items relate to the construct and each other. This led eventually to the development of modern test theory (MTT) and latent variable modeling (LVM), which are now often used as evidence for validity based on the internal structure of a scale. Although LVMs offer many advantages over CTT, they are much more computationally intensive and as such, it took a while for them to become accessible (relatedly, this type of validity evidence was not formally included in the *Standards* until the 4th edition was released in 1999). Meanwhile, the idea of construct validity was introduced with the first edition of the *Standards* in 1954, which made a distinction between the property of interest and the test designed to measure it, as well as the criterion(s) with which it was hypothesized to relate (APA et al., 1954). The shift towards construct validity was more theoretical than analytical, as “it calls for no new approach” in validation procedures (Cronbach & Meehl, 1955, p. 282), instead focusing on multiple sources of evidence in defense of the *interpretation* of test results (Slaney, 2017). However, it did align quite well with LVM given that latent variable models postulate that an unobservable latent variable is causally related to items designed to reflect it.

Since the 4th edition of the *Standards* in 1999, there have been five types of validity evidence recommended by the three committees, two of which have not yet been discussed: evidence based on the test content, response process, internal structure, relationships with external criteria, and the consequences of testing (AERA et al., 1999). Evidence of validity for the response process refers to the “fit between the construct and the detailed nature of the performance or response actually engaged in by the test takers” (AERA, APA, NCME, 2014, p. 15); in other words, the cognitive processes that participants engage in when responding to an item (Embretson, 1983). Consequences of testing emphasizes the importance of fairness and ethics in testing, suggesting that test use should be included as part of validity to avoid unfair, negative consequences of testing (Messick, 1975). The fact that these two types of validity are necessary to include speaks to the inherent difficulty of measurement in the social sciences in comparison to the physical sciences, as it would not be similarly necessary or meaningful to evaluate the consequences of creating, for example, a thermometer.

At this point, I have summarized a brief history of two important concepts in psychometrics: reliability, and validity (of which five different types of evidence can be presented). They have existed formally in the literature for varying lengths of time, however, their length of time in the literature doesn’t appear to be associated with the prominence with which each type of validity evidence is featured today. For example, Cizek et al., (2008) reviewed 283 mental assessments from the 16th edition of the Mental Measurements Yearbook (MMY) and found that 76.3% reported a reliability coefficient, 67.2% reported criterion-related evidence, 58% reported results from a factor analysis, 48.4% reported general references about the test content, and 10.6% reported face validity evidence, while only 2.5% reported evidence about the consequences of testing and 1.8% reported response process evidence. Hubley et al. (2014) reviewed 50 articles from two psychological assessment journals published between 2011 – 2012, and found that 90.2% reported a reliability coefficient, 76.8% reported criterion-related evidence, and 73.2% reported evidence based on the internal structure (factor analysis or measurement invariance), while only 1.8% reported response process and 0% reported content evidence or consequences evidence. Meanwhile, Barry et al. (2014) reviewed 967 articles in health education and behavior between 2007 and 2010, and found that 49% reported a reliability coefficient, 29% reported content evidence, 26% reported internal structure evidence (mostly factor analysis), 12% reported face validity, and 7% reported criterion-referenced evidence (while response process and consequences of testing were not mentioned). Similarly, Flake (2017) reviewed 500 scales from 122 articles in social and personality psychology in 2014, and found that 75-80% reported a reliability coefficient (mostly alpha), while 2.4 - 20.9% reported results from a factor analysis model (likewise, the authors did not investigate other types of evidence of validity).

Given the results of the meta-analyses reported above, it appears that the type of psychometric evidence reported may be dictated primarily by methodology type. Coefficients of reliability and validity (reliability, internal structure, relationships to external variables) are reported far more often than qualitative types of evidence (content, response process, and consequences). One could argue that this is due to the recency with which some types were added (e.g., evidence based on the response process and consequences were introduced last) but evidence based on the test content was one of the first types of validity evidence, introduced back in 1921, and it is scarcely found relative to its quantitative counterparts. Meanwhile, evidence based on the internal structure was formally introduced included in the *Standards* in 1999, yet factor analysis results are commonly found in the literature (relatively speaking). Slaney (2017) points out that once psychometric software became commercially available, factor analyses began to appear more often in scale validation articles. The contemporary literature on validity theory states that all types of validity evidence are equally important and that the type presented should depend on the uses and interpretations of test scores, but when it comes to reporting standards, some types are clearly prioritized over others. Michell's (2003) quantitative imperative appears to extend beyond how psychologists conceptualize the structure of psychological properties – it also dictates the type of evidence of validity that psychologists present.

In addition to the quantitative imperative, it is also possible that quantitative evidence is reported more often because it is simple and concise, and much faster to collect and synthesize than qualitative data. This is particularly important in an academic culture characterized by "publish or perish". It is also possible that emphasis on quantification is circular, i.e., it is what is seen most often and therefore it is expected to be included (e.g., some reviewers might gatekeep publishing by requiring a reliability coefficient). Further, it is worth mentioning that there do not appear to be many clear standards on *how* to investigate or report qualitative evidence. For example, when the participant's response process is mentioned, it will often be in passing in a paragraph (e.g., "we conducted several cognitive interviews with undergraduate students to determine if the items should be revised").

In the three articles that comprise this dissertation, I present some meaningful paths forward that will hopefully alleviate some of the predicaments highlighted above.

Paper One

Even though there are several established methods that can be used to present evidence of the response process, this type of validity is rarely reported. Verbal probes such as think alouds or cognitive interviews are some of the most popular approaches and are applicable for items on both academic and psychological assessments (Castillo-Díaz & Padilla, 2013; Gehlbach & Brinkworth, 2011; Leighton, 2017; Lundmann & Villadsen, 2016; Messick, 1995; Padilla & Benítez, 2014; Wolf et al., 2022). Other methods include capturing the amount of time a participant spends responding to an item, i.e., their response time (Li et al., 2017), video-ethnography (Maddox, 2017), eye tracking (Maddox et al., 2018), and log data (Oranje et al., 2017). However, none of these methods offer a clear or concise way to report results and demonstrate improvements in item interpretability.

In my first paper, I introduce a method called the Response Process Evaluation (RPE) method that I developed with Ann Taves over the last four years. We used the RPE method to develop and refine items for a scale she developed with several other colleagues called the Inventory of Non-Ordinary Experiences (INOE). We briefly introduced this method with Andy Maul in a chapter of the second edition of the *Handbook of Research Methods in the Study of Religion* (Wolf et al., 2022). That article was written when we first began to try out the method on a small test sample and we have learned a lot since then.

Briefly, the INOE is a dichotomous (yes/no) scale that captures whether people have had a variety of non-ordinary experiences (e.g., one of our items reads "I have had an experience in which it seemed as if I left my physical body”). Subsequently, participants are asked a handful of “appraisal” items about that experience (e.g., “do you consider this experience spiritual or religious?” and “do you think science can explain how this experience happened?”). The appraisal items all have a close-ended response format (some are ordered on a Likert scale, most are distinct, unordered categories). This deviates from traditional scales in the social sciences in which case researchers often include multiple items on a scale to query a hypothesized property, such as depression or well-being. As such, this scale could not be validated using traditional correlational quantitative methods, leading us to ask ourselves “how can we present evidence that participants understand the items as intended?”. This was especially complicated since the team wanted to make cross-cultural comparisons between Hindi-speaking Indians and English-speaking Americans, necessitating that we validate this survey simultaneously in two different countries and languages (making traditional qualitative methods such as cognitive interviews more challenging).

To that end, we had the unique opportunity to develop the RPE method. The RPE method is a standardized framework for pretesting multiple versions of survey items and generating individual item level validity reports. It turns cognitive interviews into open-ended surveys through the use of web probing, enabling researchers to collect data on item interpretability from a larger sample, revise items that do not appear to be understood as intended, and test the item revisions in a new sample from the same population. The RPE method stands out from other established methods of pretesting because of its unique reporting format. Specifically, the item validation reports detail the intended interpretation of each item, the population it was validated on, the percent of participants that interpreted the item as intended, and any common misinterpretations to be cautious of. This not only provides critical information necessary to properly interpret data, but also gives other researchers the information they need to determine if they can “borrow” an item for use in their own scale.

In this paper, I introduce the RPE method, demonstrate it using an INOE item, present the reporting format, and make an argument for when and why it is important to present response process evidence in support of scale use.

Paper Two

While factor analysis is commonly reported as evidence of validity based on the internal structure, there is a substantial debate in the literature as to how the fit of these models should be evaluated (Hayduk et al., 2007; Millsap, 2007; Mulaik, 2007). Though there is disagreement as to the value of approximate fit indices in assessing model fit, there is generally widespread agreement that the cutoff values currently employed are inappropriate for most models. The existing cutoffs are derived from a simulation study in 1999, which relied on a three-factor model with 15 items, each with loadings ranging from .7 to .8 (Hu & Bentler, 1999). A substantial body of literature has demonstrated these cutoffs do not generalize beyond the conditions sampled in this study, yet they are commonly used in all model subspaces (Hancock & Mueller, 2011; Heene et al., 2011; Marsh et al., 2004; McNeish et al., 2018; Saris et al., 2009). Given that the fit of a factor model is one of the most reported types of evidence of validity, there is a need to revise these cutoff values so that they will be accurate for all model subspaces.

To this end, I spent the last two years working with Dan McNeish to create a new method to compute model fit cutoffs that are tailored to the user’s individual model. We call these simulation based cutoff values dynamic fit index (DFI) cutoffs (McNeish & Wolf, 2021). Our first published article demonstrated the viability of this approach and included a brief tutorial about how to use it. With this paper, we also released a point-and-click web-based Shiny app to make this method accessible for applied researchers. It is fairly easy to use since it only requires that people upload their model statement with standardized loadings and their sample size. However, recent feedback has suggested that a tutorial paper geared towards applied researchers would be beneficial, especially when it comes to interpreting the results from the app.

Thus, my second paper is a tutorial paper for the Shiny App (first conceived in former committee member Allison Horst's class two years ago). This tutorial will follow the "ten questions" format used in Karen Nylund-Gibson's tutorial paper on latent class analysis (LCA). Specifically, I answer the following questions:

1. What is CFA and why do social scientists use it?
2. What are the different types of model fit?
3. What kinds of cutoffs do people currently use?
4. Why should I use DFI cutoffs instead?
5. How do I use DFI cutoffs?
6. What are the different levels?
7. Which level should I use?
8. How do I interpret DFI cutoffs?
9. What do the plots mean?
10. How do I include DFI cutoffs in a manuscript?
11. What does NONE mean, and what do I do if I see it?

12. What should I do if DFI cutoffs don't exist for my model type?

We plan to make this article freely available as a white paper on our website, www.dynamicfit.app. Currently, the only models that the app can generate cutoff value for are single-level CFA models with continuous outcomes estimated using maximum likelihood estimation. As we expand the website to include cutoff values for more model types, we will also update the tutorial article.

Paper Three

In my third paper, I argue that psychology has over-relied on quantitative evidence of validity over the last 100 years, resulting in survey instruments that are motivated more by desirable psychometric properties than theory and qualitative inquiry. I begin by summarizing the problems with over-generalizing Hu and Bentler's (1999) cutoffs and introducing the DFI cutoffs from Paper 2. Next, I conduct a review of some of the most popular (i.e., oft-cited) scales in psychology, locate articles that used CFA as evidence of validity for them, and re-calculate tailored DFI cutoffs for them to determine if they still meet Hu and Bentler's threshold for model fit once the cutoff vales are tailored to that particular scale.

The goal here is not to convince researchers to use DFI cutoffs but rather to encourage researchers to not rely exclusively on quantitative evidence of validity. I also document the types of validity evidence that were presented in the introductory article for each scale, reporting whether they presented each of the five types of validity evidence recommended by the *Standards*. Additionally, I note if each survey author reported reliability coefficients, a definition of the measurand, the intended uses of the scale, and instructions of how to score the survey instrument. I end this section by summarizing the implications of these findings.

The second half of this paper is a longer discussion about validity and validation in psychology. I talk about the fact that validity and validation have evolved substantially since the late 1990s, but the validation practices from the 1900s still continue to dominate modern scale construction and validation efforts (Borsboom, 2006; Zumbo & Chan, 2014). I review several articles that show why quantitative evidence of validity is insufficient. For example, Maul (2017) demonstrates that meaningless items can yield a coherent factor structure, while Hayduk (2014) demonstrates that misspecified models can still have non-significant chi-square values. As such, even if a model does fit the data well, it is difficult to conclude that it is valid for a particular purpose without multiple sources of evidence of validity. I also summarize several meta-analyses that documented the types of validity evidence typically presented in validation articles (e.g., Barry et al., 2014; Cizek et al., 2008; Flake et al., 2017; Hubley et al., 2014), highlighting that quantitative evidence is presented far more often than qualitative evidence (Chinni & Hubley, 2014).

Having demonstrated that quantitative evidence is insufficient and can lead researchers to draw incorrect conclusions about the validity of their survey instruments, I briefly introduce the RPE method from Paper 1 and describe how it can be used to test item interpretability and invite the population of interest to contribute to the construction of scales. I finish the article by discussing Michell’s (2003, 2012) quantitative imperative and urging psychologists to rely on theory when constructing scales.

Conclusion

I hope that these three papers will contribute to an improvement in validation practices in the social sciences by introducing user-friendly and practical approaches to collect and report validity evidence. I would not assume that using these methods would enable a researcher to claim that they have successfully measured a psychological attribute, but I hope that using these techniques will give audiences more confidence in the use of psychological assessments and in the interpretation of the results drawn from them. Further, I hope to free psychologists from the pressure of the quantitative imperative and encourage them to prioritize theory and rigorous inquiry over routine and ease.

References

AERA, APA, NCME. (2014). Validity. In *Standards for Educational and Psychological Testing* (pp. 11–31). American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2, Pt. 2), 1–38.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and Reliability Reporting Practices in the Field of Health Education and Behavior: A Review of Seven Journals. *Health Education & Behavior*, *41*(1), 12–18. https://doi.org/10.1177/1090198113483139

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

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium. XIV. *Journal of Educational Psychology*, *12*, 271–275.

Castillo-Díaz, M., & Padilla, J.-L. (2013). How Cognitive Interviewing can Provide Validity Evidence of the Response Processes to Scale Items. *Social Indicators Research*, *114*(3), 963–975. https://doi.org/10.1007/s11205-012-0184-8

Chan, E. K. H. (2014). Standards and Guidelines for Validation Practices: Development and Evaluation of Measurement Instruments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9

Chinni, M. L., & Hubley, A. M. (2014). A research synthesis of validation practices used to evaluate the Satisfaction with Life Scale (SWLS). In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral and Health Sciences* (pp. 229–241).

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement*, *68*(3), 397–412. https://doi.org/10.1177/0013164407310130

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

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

Gehlbach, H., & Brinkworth, M. E. (2011). Measure Twice, Cut down Error: A Process for Enhancing the Validity of Survey Scales. *Review of General Psychology, 15*(4), 380–387. https://doi.org/10.1037/a0025704

Hancock, G. R., & Mueller, R. O. (2011). The Reliability Paradox in Assessing Structural Relations Within Covariance Structure Models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Hayduk, L. (2014). Seeing Perfectly Fitting Factor Models That Are Causally Misspecified: Understanding That Close-Fitting Models Can Be Worse. *Educational and Psychological Measurement, 74*(6), 905–926. https://doi.org/10.1177/0013164414527449

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! Testing! One, two, three – Testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841–850. https://doi.org/10.1016/j.paid.2006.10.001

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*(3), 319–336. https://doi.org/10.1037/a0024917

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of Validation Practices in Two Assessment Journals: Psychological Assessment and the European Journal of Psychological Assessment. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 193–213). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_11

Kane, M. (2013). The Argument-Based Approach to Validation. *School Psychology Review*, *42*(4), 448–457. https://doi.org/10.1080/02796015.2013.12087465

Leighton, J. P. (2017). Collecting and Analyzing Verbal Response Process Data in the Service of Interpretive and Validity Arguments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of Score Meaning Using Examinee Response Processes for the Next Generation of Assessments: The Use of Response Processes* (pp. 25–38). Routledge.

Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response Time Data as Validity Evidence: Has It Lived Up To Its Promise and, If Not, What Would It Take to Do So. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 159–178). Springer.

Lundmann, L., & Villadsen, J. W. (2016). Qualitative variations in personality inventories: Subjective understandings of items in a personality inventory. *Qualitative Research in Psychology*, *13*(2), 166–187. https://doi.org/10.1080/14780887.2015.1134737

Maddox, B. (2017). Talk and gesture as process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 113–127. https://doi.org/10.1080/15366367.2017.1392821

Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G., & Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment. *European Journal of Psychology of Education*, *33*(3), 543–558. https://doi.org/10.1007/s10212-018-0380-2

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

Maul, A. (2018). Validity. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. https://doi.org/10.1037/met0000425

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*(10), 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, *50*(9), 741–749.

Michell, J. (2003). The Quantitative Imperative: Positivism, Naive Realism and the Place of Qualitative Methods in Psychology. *Theory & Psychology*, *13*(1), 5–31. https://doi.org/10.1177/0959354303013001758

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Frontiers in Psychology*, *3*, 1–8. https://doi.org/10.3389/fpsyg.2012.00261

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

Mulaik, S. (2007). There is a place for approximate fit in structural equation modelling. *Personality and Individual Differences*, *42*(5), 883–891. https://doi.org/10.1016/j.paid.2006.10.024

Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, Analyzing, and Interpreting Response Time, Eye-Tracking, and Log Data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of Score Meaning for the Next Generation of Assessments* (1st ed., pp. 39–51). Routledge. https://doi.org/10.4324/9781315708591-4

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26.1*, 136–144. https://doi.org/10.7334/psicothema2013.259

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Slaney, K. (2017). *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Spearman, C. (1904a). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–293. https://doi.org/10.2307/1412107

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/10.2307/1412159

Thurstone, L. L. (1931). Multiple Factor Analysis. *Psychological Review*, *38*, 406–427.

Wolf, M. G., Ihm, E., Maul, A., & Taves, A. (2022). Survey Item Validation. In S. Engler & M. Stausberg (Eds.), *Handbook of Research Methods in the Study of Religion* (2nd ed., pp. 612–624). Routledge.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9

Paper 1: The Response Process Evaluation Method

Introduction

Testing the coherence and relevance of survey items is commonly recommended in guidelines about scale or questionnaire construction (AERA, APA, NCME, 2014b; Gehlbach & Brinkworth, 2011; Krosnick, 1999; Presser, Couper, et al., 2004; Wilson, 2005). This can be accomplished by collecting evidence to answer questions like "was this survey item understood as intended?", "what cognitive processes are participants using to respond to items?", and "do these items function appropriately and fit the hypothesized statistical model?". Some of these questions are best answered using qualitative methods, such as cognitive interviewing, while others are better suited to quantitative approaches such as latent variable modeling (Peterson et al., 2017; Presser, Rothgeb, et al., 2004; Willis, 2004). The practice of collecting such evidence is known as scale validation, which is undertaken to assess the adequacy or appropriateness of the use of a survey for a particular purpose (Kane, 2013; Maul, 2018; Sireci, 2007).

According to the *Standards for Educational and Psychological Testing* (2014a), validation is an on-going process for which five different types of validity evidence can be presented dependent upon the intended interpretations of the survey results: evidence based on the test content, response process, internal structure, relationship to other variables, and the consequences of testing. The goal is to demonstrate that the scale is scientifically sound and that the inferences generated from its use are trustworthy (Sireci, 2007). Psychologists constructing scales often prioritize validity evidence that is more quantitative in nature, such as latent variable modeling or regression coefficients, which are typically used to evaluate the internal structure of a scale or its relationship to external variables (Cizek et al., 2008; Flake et al., 2017; Hubley et al., 2014). The emphasis on quantitative evidence may be caused in part by what Michell (2003) describes as the “quantitative imperative”: the idea that quantity is necessary for measurement and measurement is necessary for psychology to be considered scientific (Slaney, 2017). While each type of validity evidence may not be warranted for every validation effort, omitting response process evidence entirely essentially places researchers in a position in which they must defend the validity of a scale without presenting evidence that participants understand survey items as intended. It is difficult (if not impossible; Sireci, 2007) to conclude that scales measure what they intend to measure, but it is even more challenging to make this claim if there is not congruence in item interpretation and meaning between scale creators and scale respondents.

To test item comprehension, researchers usually collect evidence of validity based on the participant’s response process. The participant’s response process is the cognitive process that an individual engages in when responding to an item on an assessment (Padilla & Benítez, 2014). Evidence of validity for the response process refers to the “fit between the construct and the detailed nature of the performance or response actually engaged in by the test takers” (AERA, APA, NCME, 2014, p. 15). Pretesting questionnaire items before they are administered can reduce misinterpretations, improve item clarity, and help ensure that survey items function as intended (Hilton, 2017; Willis, 2005). This evidence has been described as essential (Launeanu & Hubley, 2017) or as validity in its own right (Wilson, 2005). Borsboom et al. (2004) go a step further and argue that a test can *only* be valid if variability in the attribute of interest causes variation in assessment scores (i.e., that the response process is all of validity theory). In addition, it is also common test item clarity by having experts or graduate students review items for appropriateness and interpretability. While this is an important step in item development, it does not ensure that the population of interest will use the intended cognitive processes or agree that the item reflects the property of interest (Peterson et al., 2017).

Response process evidence is often collected using verbal probing methods such as cognitive interviews or think alouds (Castillo-Díaz & Padilla, 2013; Gehlbach & Brinkworth, 2011; Leighton, 2017; Lundmann & Villadsen, 2016; Messick, 1995; Padilla & Benítez, 2014). Cognitive interviews are a type of semi-structured interview protocol that consist of questions or "probes" which can be used to elicit information about the cognitive process that the participant uses to respond to survey items. Participants are encouraged to either "think aloud" while they respond to a survey item, or answer questions retrospectively about their thought process after they respond (Priede & Farrall, 2011). These methods can be used to gather information about the "question-and-answer" process model (Tourangeau, 1984), i.e., the phases that participants transition through when responding to a survey item (comprehension, recall, judgment, and response). Participants typically begin by interpreting the item, retrieving the information necessary to respond to it, evaluating the information retrieved given their comprehension of the item or task, and selecting a corresponding response option (Padilla & Benítez, 2014). General probes might include questions such as "can you repeat that item in your own words?" or "what were you thinking about when you selected that response option?". Probes can be tailored based on anticipated misinterpretations or areas of concern within and across cultures (Peterson et al., 2017).

While pretesting items is an invaluable practice, conducting cognitive interviews and summarizing the resulting data can prove lengthy and difficult (Castillo-Díaz & Padilla, 2013; Launeanu & Hubley, 2017). Web probing has recently been introduced as a rapid, survey-based approach to cognitive interviewing in which response process evidence is collected electronically through open-ended responses (Behr et al., 2017; Edgar et al., 2016; Fowler & Willis, 2020). The probes used in web probing are very similar (if not identical) to those used in cognitive interviews; the main difference is the medium in which data is collected. Cognitive interviews are typically conducted in person and thus create the opportunity for the interviewer and interviewee to engage in a dialogue about the participant's cognitive process in which the interviewer can ask follow-up probes if the interviewee's response is unclear or brings up an interesting point. Web probes use the same probing questions but in an open-ended survey format, typically collecting data online using a crowdsourcing platform such as MTurk or Prolific. While there is no opportunity to ask instantaneous follow up questions, the interviewing process is more standardized and data can be generated much more quickly and easily than with cognitive interviewing (Edgar et al., 2016; Fowler & Willis, 2020). Administering web probes online instead of in a local lab also makes it possible to reach a wider participant pool, which could help researchers target specific populations and understand how item interpretations may vary across different demographics (Edgar et al., 2016). While the results of cognitive interviews often come from small samples and are not necessarily intended to generalize to a larger sample, web probing provides the opportunity to generate a much larger sample, which means that study inferences from web probes are less likely to suffer from small sample over-generalizations and makes it more likely that errors in interpretation will be uncovered (Behr et al., 2017; Edgar et al., 2016; Meitinger & Behr, 2016).

After pretesting items, it is common to find that some items were misinterpreted and need to be revised before being administered. It is desirable to test these revised items in a new sample to ensure that modifications improved the interpretability of the item (Peterson et al., 2017; Ryan et al., 2012). Although cognitive interviews and web probing are useful for testing the interpretability of one version of an item, they do not offer a clear or simple way to iteratively test multiple versions of an item that was not understood as intended. This makes it difficult to prove that any revisions have made the item clearer or reduced the frequency of misunderstandings (Willis, 2005). Further, there are no reporting frameworks in which researchers can document important features of item-level validity, such as the intended interpretation of an item, the rationale for its inclusion in a scale or questionnaire, common misinterpretations, the characteristics of the population in which it was tested, and the percent of respondents that understood the item as intended. Indeed, it is often noted that existing pretesting protocols are often too vague to reproduce reliably and that they lack the necessary guidelines necessary to be implemented successfully (Hilton, 2017; Presser, Couper, et al., 2004). Including this kind of information might increase the transparency and replicability of the psychological sciences and make it easier for researchers to confidently “borrow” items from established scales.

The Response Process Evaluation Method

The Response Process Evaluation (RPE) method is a standardized framework for pretesting multiple versions of survey items and generating individual item level validity reports (Wolf et al., 2022). It turns cognitive interviews into meta-surveys, employing web probes to develop and validate items in an iterative fashion[2]. When used on a crowdsourcing platform, researchers can quickly gain insights into the interpretability of survey items, make revisions when necessary, and retest item interpretability with a new sample of respondents. This process is repeated sequentially until the final version of each item is constructed (or the item is removed from the scale due to non-convergence in shared meaning across participants). The result is a set of item validation reports that detail the intended interpretation of each item, the population it was validated on, the percent of participants that interpreted the item as intended, and any common misinterpretations to be cautious of.

Like cognitive interviews and web probes, the RPE method uses probes that prompt participants to explain two aspects of their response process: their interpretation of the item and the rationale for the response option they selected. These probes are designed to elicit evidence that enables researchers to evaluate the extent to which the item was understood as intended or might warrant a testable revision. The probes are administered in a meta-survey in which participants respond to open-ended questions about a subset of survey items rather than completing the entire survey in full. For example, participants might be asked to restate the item in their own words (paraphrase probe), define a key word or phrase in the item (comprehension probe), and explain why they selected a particular response option (category-selection probe; Behr et al., 2017; Willis, 2005).

Participants provide written responses to these probes which are subsequently evaluated by researchers and coded as either “understood”, “likely understood”, “likely not understood”, “not understood”, or “not enough information”. Each response is coded

[2] Data collection is not restricted to the web. Meta-surveys can also be administered using pen and paper.

holistically, combining information from all the probes to make an overall decision for each participant and each item. Data is collected in *batches* of five participants and coded by at least two trained subject matter experts that are clear as to the intended interpretation of each item and the intended use of the scale. After a batch of responses is coded, researchers meet to compare their codes and discuss how well each item appears to be functioning. Survey items that are coded as "understood" or "likely understood" by the researchers are readministered to another batch of five participants to collect more data. Items that are coded as "likely not understood" or "not understood" are either revised and then readministered to a new batch of five participants, retested to collect more data, or removed from the survey. If an item has an unusually high number of "not enough information" codes, additional probes should be added to the next batch of data collection to reduce ambiguity. This process is repeated iteratively until the final version of the item has been evaluated an adequate number of times. In our experience, we began to see less variation and lower return on investment after twenty responses to the final version of the item; this number is slightly higher than the five to fifteen participants recommended in cognitive interviews (see, e.g., Peterson et al., 2017). Thus, we tentatively recommend that the final version of each item be evaluated twenty times and be considered validated[3] when it is coded as "understood" or "likely understood" at least 80% of the time.

If items are written to reflect a particular content domain (e.g., math ability or depression), it may be helpful to include a probe that queries the extent to which participants agree that (1) the survey item is relevant for that content domain and (2) that the content domain has been adequately covered by the set of survey items (Peterson et al., 2017).

[3] The "validity" of the item could change over time due to cultural shifts or new populations.

Querying this from participants essentially invite them to contribute to the development of the scale by producing evidence of validity based on the test content (e.g., the relevance of the test items from the perspective of the participants; Dumas, 2008). This type of validity evidence is often gathered by consulting experts or relevant literature (AERA, APA, NCME, 2014a; Lissitz, 2009; Sireci, 1998) but this does not guarantee that the items are relevant to participants, especially when researchers seek to measure psychological properties that are more ontologically subjective or culturally specific. If researchers intend to use the scale to make cross-cultural comparisons or administer the survey to participants from a different culture, then the entire RPE process should be repeated using participants from that group to ensure it is cross-culturally sound (for more information, see Wolf et al., 2022 and Taves et al., in preparation).

We developed the framework of the RPE method to create a well-documented, evidence-based approach to item development and incorporate participants into the item revision process[4]. The strength of the method lies in (1) its unique ability to revise and retest items iteratively until the final version of an item is reached, and (2) create user-friendly item level validity reports that make it easier to judge the adequacy of a survey item for a particular purpose and borrow items for use in other studies. In the next section, we will walk through the validation process for one survey item and witness how feedback from the population of interest helps revise the item and clarify the concept of interest.

Using the RPE Method

[4] The framework used in the RPE method could be applied for other purposes, such as testing the meaning of a word or phrase across cultures.

To demonstrate the RPE method, we will use an item from the Inventory of Non-Ordinary Experiences (INOE) that was recently validated for English-speaking Americans using the crowdsourcing platform MTurk. The INOE is comprised of 40 close-ended "experience" items in which participants respond "yes" or "no" to whether they have had a specific experience, e.g., "I have had the experience of being aware that I was dreaming while asleep". If participants respond affirmatively, they are asked a series of close-ended follow-up questions about that experience, e.g., "Overall, how much of an impact has this experience had on your life?". The items are intended to be understood by a lay population, and the instrument is intended to be used to compare the frequency and perceived origin of non-ordinary experiences across two cultural groups: English-speaking Americans and Hindi-speaking Indians. In practice, this item was validated concurrently for both English-speaking Americans and Hindi-speaking Indians with a cross-cultural team of researchers. For simplicity, we elect to demonstrate the validation process only for English-speaking Americans, but the process can be expanded to validate items simultaneously in multiple languages. Readers that are interested in detail about the cross-cultural validation process are referred to Taves et al. (in preparation).

Many of the original items on the INOE were adapted from or inspired by existing scales in the literature about religious, paranormal, or psychotic experiences because these experiences are often considered to be extraordinary, unusual, or rare. The INOE research team wanted to include a survey item about non-ordinary experiences of love and compassion because these two emotions are often cultivated by religious traditions (e.g., "God's love"). Love and compassion were combined into one item because the two terms are often combined colloquially, and because the research team felt the terms were similar. The first *iteration* of the survey item was “I have had an experience of love or compassion that stood out from all other such experiences” (Table 1). Note that participants should only respond affirmatively if the experience stood out (i.e., was non-ordinary or unique). Also note that a clear definition of “love and compassion” and an “intended interpretation” was not included when we first began the validation process for this item.

Table 1. The first iteration of the love and compassion item and corresponding instructions.

Item Instructions	Please indicate whether or not you have had each kind of experience, by selecting ’Yes’ or ’No’. Only select ’yes’ if you can remember at least one specific experience that stands out.
Survey Item	I have had an experience of love or compassion that stood out from all other such experiences

We used five probes to capture each participant’s response process (Table 2). To better replicate the survey taking experience, participants first responded to the item itself (response probe). Next, they were asked to paraphrase the entire item in their own words (paraphrase probe) followed by defining the key terms in the item (comprehension probe). We found that it was necessary to include both the paraphrase and comprehension probes for this item because without the comprehension probe, participants would sometimes only paraphrase the “stand out” part of the item and mistakenly omit the meaning of love and compassion. These two probes would sometimes yield redundant information if participants defined love and compassion as part of their response to the paraphrase probe. There were two versions of the category-selection probe; the version each participant saw was triggered by their response to the survey item. If they selected “yes”, they were asked to briefly describe the experience; if they selected “no”, they were asked to give an example of such an experience. The last probe they responded to was an optional catch-all probe in which they were invited to share any other feedback they had. These probes were developed using trial-

and-error in which the order and wording of each probe was modified until it elicited the information necessary to determine if the item was understood as intended.

Table 2. The probes used in the meta-survey.

Probe Type	**Probe Wording**
Response	If these were the response options, which would you select?
Comprehension	In your own words, how would you explain "love or compassion"?
Paraphrase	In your own words, what do you think this item means?
Category-selection [Yes]	Briefly describe your experience of love or compassion that stood out from all other such experiences.
Category-selection [No]	Please give an example of such an experience even though you have not had one.
Catch-all	Is there anything you don't understand or would change? If so, what?

We estimated that responding to all the probes for one item would take approximately 3 minutes and typically administered meta-surveys with 3 items per survey to reduce participant burn out (i.e., each survey would take approximately 9 minutes to complete) and paid minimum wage proportional to the amount of time the survey was expected to take. Meta-survey length and compensation is an important consideration with surveys that are comprised of mostly open-ended items because open-ended survey items have a much higher cognitive burden than close-ended survey items (Dillman et al., 2014; Tourangeau et al., 2000). Responses were collected in batches of five, meaning that the research team coded responses from five participants at once. We kept all the items and probe responses in the meta-survey within the same coding file so that we could see if a participant consistently gave low-effort responses or unusual throughout the meta-survey (at which point we would consider removing them; Moss, 2018). We found that approximately 5% of the American participants on MTurk gave responses that were unusable.

A sample of responses from the first two batches of data for the first iteration of the survey item are presented in Table 3. Three of the participants selected “yes” and one participant selected “no” when responding to the item “I have had an experience of love or compassion that stood out from all other such experiences”. When reading through the first participant’s responses, we noticed that their responses to the comprehension and paraphrase probes seemed in line with what we intended but their description of themselves as a “compassionate person” with “intense feelings of love *all of the time*” when responding to the category-selection probe indicated that they did not recall a “stand out” or “unique” experience when responding. This led us to code their overall response as “not understood” even though their responses to the comprehension and paraphrase probes were good. If we had only used one probe to pretest this survey item, it would have been difficult to evaluate the participant’s response process and determine if they understood the item as intended.

Table 3. A sample of responses from the first batch of data for the first iteration of the love and compassion item.

	Response	Comprehension	Paraphrase	Category-Selection
1	Yes	love is a strong feeling of affection compassion is sympathy and concern for others	Have I ever felt strong concern for others or intense affection for someone which stands out?	I have intense feelings of love all of the time - for my children, for my husband. I also am a compassionate person.
2	No	A warm emotion you feel deep, straight in your heart	If we felt near we had a very strong experience in which and loving towards someone	I don't remember of any experience particularly, but I'm lucky to have a loving fiancee, family, and friends
3	Yes	intense feeling of affection or sympathy	Asking if I've ever experienced love or compassion.	After my dog died, I felt compassion for animals and started eating meat 1-2 times of months only.

4	Yes	Sympathetic pity or concern for others.	Having pity or concern for others.	When my dad died I had an overwhelming sense of compassion and love for my younger brothers.

The second participant was coded as "likely understood" because while their response to the comprehension probe was vague, their response to the category-selection probe was exactly in line with what we intended (e.g., they selected "No" when responding to the item because they didn't remember a specific stand out moment despite feeling loving feelings for people close to them). Comparing the first and second participant yields a classic case of multifinality, e.g., the idea that participants can use the *same* cognitive processes when responding to an item but select *different* response options (Lundmann & Villadsen, 2016). Both participants described themselves as having loving relationships, but the first participant used this as a justification to respond "yes" to the item while the second participant chose to respond "no". The third and fourth participant were both coded as "understood"; while they did not properly paraphrase the "stand out" nature of the item, their "yes" responses combined with a clear memory of a meaningful experience of compassion in their responses to the category-selection probe made it clear to us that they both understood the item as intended.

When discussing the responses to the first two batches of data, two things became clear: (1) compassion and love were two distinct emotions that should probably not be combined into the same item, and (2) that we needed to clearly delineate the intended interpretation of the item to ensure that coders were on the same page. The first point arose upon reading the responses and noticing that the second participant only described feelings of love while the fourth participant only described feelings of compassion. We worried that participants might mistakenly focus on either love *or* compassion when reading the item and therefore end up using very different cognitive processes when responding. The second point arose from lively debates during our team coding review meetings in which it became clear that while the definitions of love and compassion seemed self-evident, our disagreements on how the responses should be coded indicated that an unambiguous intended interpretation would streamline the validation process. Thus, we split the item into two items and added intended interpretations for both (see Table 4).

Table 4. Intended interpretation of each survey item.

Item	Intended Interpretation
Love	Love is a deep feeling of affection and attachment, which includes but is not limited to romantic feelings (OED Online, n.d.). Religious traditions often cultivate this feeling, so the feeling can be of any duration. It does, however, need to "stand out" from other such experiences. This feeling is likely most common in relation to people, but also may be felt in relation to animals or in relation to various contexts or without any apparent precipitating cause.
Compassion	Compassion is a “sympathetic pity and concern for the sufferings or misfortunes of others” (OED Online, n.d.). This should only be endorsed when an individual can recall a specific experience of compassion that they felt for others that “stands out” to them; it should not be interpreted as a reflection of how compassionate one believes themselves to be.

For the sake of brevity, we continue the demonstration of the RPE method by presenting data only from the second and final *iteration* of the compassion item. In the second iteration, we decided to simply remove the word “love” and otherwise retain the same structure for the item stem. Thus, the second iteration of the compassion item read “I have had an experience of compassion that stood out from all other such experiences”. A sample of responses to the second iteration of the compassion item can be seen in Table 5. When coding this item, we noticed that some participants were recalling or giving examples of specific experiences in which they witnessed or felt compassion *from* others rather than times they personally felt compassion *for* others, which was not in line with what we intended. As such, participant one and participant three's responses were coded as "not understood". Participant two's response was coded as "not enough information" because it was vague and we were unable to come to an agreement as to whether they recalled a specific experience. Participant four's response was coded as "understood" because their responses to all three probes were in line with what we imaged when writing the item. The responses to the probes for the first and second iteration of this item made it clear that we needed to further revise the item to (1) clarify we wanted a feeling of compassion *for* others, and (2) better highlight the "stand out" nature of the experience.

Table 5. Responses to the probes for the second iteration of the compassion item.

	Response	Comprehension	Paraphrase	Category-Selection
1	No	An experience of compassion is an instance where I've felt or have seen others exhibit an instance of compassion that was distinct from ordinary sorts of compassion.	It could be compassion I've demonstrated, compassion someone demonstrated toward me, or compassion a third party demonstrated that I saw.	A homeless man giving an even more desperate person their remaining cash out of charity.
2	Yes	I feel bad for people that are less fortunate because nobody should live like that.	It means having compassion for people.	When people didn't have food to eat and went to sleep hungry.
3	Yes	Something where someone showed compassion for another person	Have you an experience where you or someone else showed compassion that was above and beyond any that you had felt before.	We lost our house to foreclosure and couldn't find as place to rent because of our credit. One home owner to compassion on us by renting there house to us without even looking at our

				credit because they knew our situation
4	Yes	i felt bad or sorry for someone or people because they were in a bad situation, or something bad had recently happened to them.	think of a time when you felt compassion for others	seeing that young man and his two year old girl washed up on a shore dead because they had to escape their country because of the u.s. starting wars where they lived

The compassion item was revised three more times to address the issues above before we reached the final iteration. The sixth and final iteration of the compassion item read: “I can recall a specific experience in which I felt compassion for the suffering of others (human or nonhuman) that stood out from all other such experiences” (emphases are part of the item). We found that it was necessary to underline the “stand out” part of the item stem to increase the probability that respondents would notice it. A sample of responses to the probes are presented in Table 6. All four responses were coded as “understood” because it was clear that each participant understood that the experience must be uniquely memorable and that the emotion must be one that they felt towards others. Note that even if the participant did not properly paraphrase the “stand out” part of the item, their response to the category-selection probe made it possible to determine that they were thinking of a discrete, meaningful experience.

Table 6. Responses to the probes for the final iteration of the compassion item.

	Response	Comprehension	Paraphrase	Category-Selection
1	Yes	Did you feel emotional pain in sympathy to someone else	Is there a situation that was so vivid it stands out to me in regards to my emotional response?	Watching my grandmother deal with the loss of her brother

2	Yes	Feeling sadness or empathy when something bad or painful happens to a human or an animal.	Was some event so distinct that it was exceptional and memorable.	My cat died and he really seemed to be in pain (he died at the vet). I still think about this experience at least once a month and it brings me to tears, even after been 9 years.
3	No	wanting someone who is suffering to no longer suffer	a time you remember something more than any other	Seeing a homeless family, not just a person, but a mother or father and their child or children.
4	Yes	Being able to understand the hardship of the people suffering.	Imaging yourself in those peoples shoes.	On 9/11 the families who lost loved ones when I saw interviews on TV, it was heart breaking.

The validation process for the sixth and final item started out the same as the others: with a batch of five respondents. After each participant's response was coded as "understood" or "likely understood", we readministered the same item to a new batch of five participants in order to collect more data and determine if the trend of correct interpretations replicated in a new sample. This process was repeated until we collected data from 20 participants. In the end, 19/20 participants understood the item as intended (95%) with one participant coded as "likely not understood". If we had administered the second iteration of this survey item without going through this extensive validation process, roughly 25% of participants would not have understood the item as intended, making the conclusions based on the analysis of their data inaccurate. Using the RPE method resulted in an item that we feel confident is interpreted by participants as we intended it to be interpreted.

Constructing Probes

There is no correct way to write a probe; a probe is correctly written when it elicits the information necessary to determine if an item was understood as intended. Given that an interviewer is not present to ask follow-up questions, it is important to use unambiguous probes that clearly instruct the participant to provide actionable information about their response process (Behr et al., 2017; Willis, 2005). For example, we quickly learned that using “How would you respond and why?” as a category selection probe was worded too generally when we received responses such as “I'd say no”, “I would respond in a positive way”, or “Sometimes, but not very frequently”. Even without the context of the item, it is clear that these responses do not provide enough information to judge if the item was misinterpreted. When we changed the probe to “Please explain why you selected this response” or “Briefly describe the specific experience in which you felt compassion for the suffering of others (human or non-human)”, we were able to get much more actionable responses that helped us evaluate the participant’s response process and determine if the item was understood as intended.

We found that multiple probes were necessary to follow the participant’s response process and make a holistic judgement as to whether the item was understood as intended by the participant. Edgar et al. (2016) compared the data quality derived from laboratory interviews with single-question web probes and found a similar result: their most actionable data came from the inclusion of unscripted follow-up probes in laboratory interviews. When designing probes for the RPE method, it is therefore important to construct them in such a way that a follow-up probe would not be needed. In pursuit of this, researchers should focus on probe clarity rather than including multiple redundant probes to avoid unnecessarily burdening or frustrating participants. Collectively, the probes should address the four phases of the cognitive process: comprehension, recall, judgment, and response (Tourangeau, 1984). This may require rewriting probes several times, revising the order of the probes, or testing them in a laboratory setting or with family and friends. It is important to remember that participants cannot get anything wrong; rather, if the item or probe is misunderstood, it is the researcher's responsibility to clarify it.

Coding Responses

Like cognitive interviews and web probing, we recommend having multiple coders for each response. Throughout the validation of the INOE items we used a minimum of two coders and a maximum of five but felt that three or four was the optimal number. The meta-survey responses for each batch of participants were distributed individually to all coders in private spreadsheets. We made sure to keep the responses for all items within the same spreadsheet so that we could evaluate participant quality and effort (e.g., we do not want to code a response as "not understood" because of a participant that consistently gave extremely low-effort responses – we instead suggest removing these participants). Researchers then place their individual codes and comments into a shared evaluation report which should also contain all participant responses to each probe and some administrative details about the meta-survey (see Table 7 for a condensed example and https://osf.io/uy8sr for the full evaluation report). Subsequently, researchers meet to discuss their codes within the shared evaluation report and form a group consensus on the overall code for each participant: "understood", "not understood", or "not enough information". Responses that are coded as "not enough information" are removed from the total count when calculating the percent of responses that were understood as intended (e.g., if we calculated the percent understood for the data in Table 7 the result would be 33%).

Table 7. A condensed example of the Compassion evaluation report for iteration #2.

Iteration	Batch	Participant	Evaluator 1	Evaluator 2	Overall	Modification
2	1	1	LN	N	N	"I can recall a specific experience in which I felt compassion for the suffering of others." +
		2	I	LU	I	
		3	N	N	N	
		4	LU	U	U	

Note. U = Understood, LU = Likely understood, LN = Likely not understood, N = Not understood, I = Not enough information.
$^{+}$ This was the third iteration of the Compassion item and was not included in this article.

Some disagreement during the coding process is to be expected; it is mostly a concern when researchers are assigning codes with opposite meanings (e.g., Researcher A assigns a code of "Understood" and Researcher B assigns a code of "Likely not understood"). In our experience, this is the result of one of three causes: (1) the coders are not on the same page about the intended interpretation of the item and uses of the scale, (2) the participant's response was ambiguous, or (3) one researcher noticed something that another researcher did not. The first can be resolved by improved communication between researchers and more training or by refining the intended interpretation, the second can be resolved by modifying existing probes, adding additional probing questions, and/or assigning that participant a final code of "not enough information", and the third requires no significant resolution as the researchers should be able to come to an agreement during the coding review. It is important to keep in mind that survey questions that require the participant to reflect, think critically, and give written responses are associated with a higher cognitive burden and that participants may not be motivated to give extremely detailed responses to every probe (this may be a cause of researcher disagreement; Dillman et al., 2014). This is another reason that we recommend using a holistic coding approach in which the final determination of understanding is made by reading through the responses to all the probes for each item. We also stress the importance of documenting the intended uses and interpretations of the scale results, the intended population for which it is to be administered, and the intended interpretation and rationale for inclusion of each item to ensure that coders can reliably evaluate the alignment between the item intent and participant's response process (Castillo-Díaz & Padilla, 2013; Peterson et al., 2017).

Producing user-friendly item validation reports is paramount to ensure that the most important information is easily accessible and actionable for readers. Unfortunately, this often means that there is not enough space to document the wording of the final item, the intended interpretation of each item, examples of participant interpretations, the population for which it was validated, the percent of participants that understood the item as intended, the intended uses of the instrument, and common misinterpretations, especially for items that are excluded from the final version of the instrument (Willis, 2005). This information can instead be placed in an appendix or made available online in an item repository such as the Open Science Framework (OSF). Alternatively, a centralized searchable item repository could be created which would act as a warehouse of items from all surveys[5]. It is expected that this might be most valuable for reviewers critiquing the instrument and validation process, as well as individuals hoping to borrow items for use in their own research. The item validation report for the Compassion item can be found at https://osf.io/q94b7.

Conclusion

While it is generally agreed that survey items should be pretested for interpretability through the collection of response process data such research is rarely published, rendering

[5] This idea arose during a conversation between the lead author, Gjalt-Jorn Peters and Danny Katz.

the extent to which it is conducted questionable (Cizek et al., 2008; Fowler & Willis, 2020; Hubley et al., 2014; Peterson et al., 2017). Current best practices for collecting such evidence, cognitive interviewing and web probing, do not have a standardized reporting framework or offer a way to iteratively test multiple versions of the same item to determine if modifications have improved item interpretability (Castillo-Díaz & Padilla, 2013; Hilton, 2017). The RPE method turns cognitive interviews into meta-surveys through the use of web probing, allowing researchers to iteratively test multiple versions of items, quantify and qualify improvements in interpretability, and creates the possibility of a searchable item repository through the generation of standardized item validation reports. Thus, the RPE method allows researchers to test if there is a shared understanding about item meaning with their target population on a large scale; a practice that might seem foundational to social science research but has thus far been unobtainable.

Some may argue that response process evidence is only necessary to present if one wanted to make an argument that an individual was using the appropriate or expected cognitive processes to respond to an item (AERA, APA, NCME, 2014a; Kane & Mislevy, 2017). This type of evidence may not always be necessary to collect. For example, in the medical field, the mechanism of action for medications is often unknown (in other words, a medication functions as intended but researchers aren't sure why it produces a therapeutic effect). This can be thought of as a "black box" in psychology; an unknown link between the participants responses to an item and their overall assessment score (Launeanu & Hubley, 2017). As such, it is reasonable that psychologists may be primarily interested in the classification accuracy of an assessment (i.e., relations to other variables) and not the processes underlying assessment responses. For example, if a placement test does a good job of properly assigning students to a class of the appropriate skill level, then it might not matter to a researcher if performance on the placement test is primarily due to the test-taker's social capital and financial resources rather than their knowledge of the content domain of the construct. However, if understanding the psychological processes that generated the data was an important element of the researcher's validity argument, then it would be their responsibility to provide evidence to refute any meaningful rival hypotheses about the factors driving test performance (Messick, 1995; Padilla & Benítez, 2014). The RPE method can help researchers test these hypotheses by demonstrating a priori how items are interpreted and the degree to which participants agree that they are representative of the domain of interest (when relevant).

Using the RPE method to test item interpretability essentially invites participants to contribute to the development of survey items. By giving participants an opportunity to provide with survey creators with feedback about the meaning of items or the construct of interest, researchers are adopting a bottom-up, collaborative approach to construct definition and item development (Launeanu & Hubley, 2017). Collaborating with the population of interest may be more critical when the property of interest is ontologically subjective and can vary significantly cross-culturally. Collecting this feedback in a meta-survey format can be advantageous because participants are less likely to suffer from social desirability bias in an anonymous setting than in a face-to-face lab setting (Krumpal, 2013; Tourangeau & Yan, 2007). Thus, participants that typically skim surveys when completing them can continue to do so, making it easier to notice when items are misread because they lack emphases or are too long or complicated (Bowling et al., 2021). Participants may also feel more comfortable giving honest feedback without someone observing them.

Pretesting items before using them to collect data is critical because there is no amount of statistical manipulation that can resolve poorly worded or poorly chosen items (Streiner & Norman, 2008). If the items are consistently misinterpreted by certain groups, then it may be discovered by analyses such as differential item functioning, tests of measurement invariance, or applications of mixture modeling, and it is possible that these items could be removed (although without qualitative inquiry, it will be unclear why these items did not function as intended or if removing them threatened the validity of the scale; Bond, 1993; Padilla & Benítez, 2014). Items that are consistently misinterpreted irrespective of group membership might be undetectable by statistical modeling, especially if there are many items that are consistently misinterpreted. Further, the use of these latent variable modeling methods necessitates that the items are written to reflect an unobservable property, which may not always be the case. The RPE method framework helps shifts the primary approach to validation in psychology away from pure quantification. While psychometric models are important, we believe that the practice of scale validation would benefit from testing the interpretability of survey items.

References

AERA, APA, NCME. (2014a). *Standards for Educational and Psychological Testing*. American Educational Research Association.

AERA, APA, NCME. (2014b). Test Design and Development. In *Standards for Educational and Psychological Testing* (pp. 75–94). American Educational Research Association.

AERA, APA, NCME. (2014c). Validity. In *Standards for Educational and Psychological Testing* (pp. 11–31). American Educational Research Association.

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive, interviewing in web surveys with the goal to assess the validity of survey questions. *Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_en_023

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277–279). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1075/z.62.13kok

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

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the Questions Ever End? Person-Level Increases in Careless Responding During Questionnaire Completion. *Organizational Research Methods*, *24*(4), 718–738. https://doi.org/10.1177/1094428120947794

Castillo-Díaz, M., & Padilla, J.-L. (2013). How Cognitive Interviewing can Provide Validity Evidence of the Response Processes to Scale Items. *Social Indicators Research, 114*(3), 963–975. https://doi.org/10.1007/s11205-012-0184-8

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement, 68*(3), 397–412. https://doi.org/10.1177/0013164407310130

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (4th edition). Wiley.

Edgar, J., Murphy, J., & Keating, M. (2016). Comparing Traditional and Crowdsourcing Methods for Pretesting Survey Questions. *SAGE Open*, *6*(4), 215824401667177. https://doi.org/10.1177/2158244016671770

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

Fowler, S., & Willis, G. B. (2020). The Practice of Cognitive Interviewing Through Web Probing. In P. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, & A. Wilmot (Eds.), *Advances in Questionnaire Design, Development, Evaluation and Testing* (1st ed., pp. 451–469). Wiley. https://doi.org/10.1002/9781119263685.ch18

Gehlbach, H., & Brinkworth, M. E. (2011). Measure Twice, Cut down Error: A Process for Enhancing the Validity of Survey Scales. *Review of General Psychology*, *15*(4), 380–387. https://doi.org/10.1037/a0025704

Hilton, C. E. (2017). The importance of pretesting questionnaires: A field research example of cognitive pretesting the Exercise referral Quality of Life Scale (ER-QLS).

International Journal of Social Research Methodology, *20*(1), 21–34. https://doi.org/10.1080/13645579.2015.1091640

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of Validation Practices in Two Assessment Journals: Psychological Assessment and the European Journal of Psychological Assessment. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 193–213). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_11

Kane, M., & Mislevy, R. (2017). Validating Score Interpretations Based on Response Processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of Score Meaning Using Examinee Response Processes for the Next Generation of Assessments: The use of the Response Process* (pp. 11–21). Routledge. https://doi.org/10.4324/9781315708591-2

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Launeanu, M., & Hubley, A. M. (2017). Some Observations on Response Processes Research and Its Future Theoretical and Methodological Directions. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 93–113). Springer.

Leighton, J. P. (2017). Collecting and Analyzing Verbal Response Process Data in the Service of Interpretive and Validity Arguments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of Score Meaning Using Examinee Response Processes for the Next Generation of Assessments: The Use of Response Processes* (pp. 25–38). Routledge.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions, and applications*. Information Age Pub.

Lundmann, L., & Villadsen, J. W. (2016). Qualitative variations in personality inventories: Subjective understandings of items in a personality inventory. *Qualitative Research in Psychology*, *13*(2), 166–187. https://doi.org/10.1080/14780887.2015.1134737

Maul, A. (2018). Validity. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Meitinger, K., & Behr, D. (2016). Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results? *Field Methods*, *28*(4), 363–380. https://doi.org/10.1177/1525822X15625866

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, *50*(9), 741–749.

Moss, A. (2018, September 18). After the Bot Scare: Understanding What's Been Happening With Data Collection on MTurk and How to Stop It. *CloudResearch*. https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26.1*, 136–144. https://doi.org/10.7334/psicothema2013.259

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for Testing and Evaluating Survey Questions. *Public Opinion Quarterly*, *68*(1), 109–130. https://doi.org/10.1093/poq/nfh008

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (Eds.). (2004). *Methods for Testing and Evaluating Survey Questionnaires*. John Wiley & Sons.

Priede, C., & Farrall, S. (2011). Comparing results from different styles of cognitive interviewing: 'Verbal probing' vs. 'thinking aloud.' *International Journal of Social Research Methodology*, *14*(4), 271–287. https://doi.org/10.1080/13645579.2010.523187

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving Survey Methods With Cognitive Interviews in Small- and Medium-Scale Evaluations. *American Journal of Evaluation*, *33*(3), 414–430. https://doi.org/10.1177/1098214012441499

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*, 299–321.

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Slaney, K. (2017). *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.).

Tourangeau, R. (1984). Cognitive science and survey methods: A cognitive perspective. In T. Jabine, M. Straf, J. Tanur, & R. Tourangeau (Eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between the Disciplines* (pp. 73–100). National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

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

Willis, G. B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design* (1st edition). SAGE Publications, Inc.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Taylor & Frances.

Wolf, M. G., Ihm, E., Maul, A., & Taves, A. (2022). Survey Item Validation. In S. Engler & M. Stausberg (Eds.), *Handbook of Research Methods in the Study of Religion* (2nd ed., pp. 612–624). Routledge.

Paper 2: Dynamic Fit Index Cutoffs: A Tutorial

Confirmatory factor analysis (CFA) is a commonly used statistical method in the social sciences. The goal of factor analysis is to reduce a set of observations down to a smaller number of dimensions, or *factors*, based on common features or patterns in the data. Although these models have been used for over a century, debate remains about how to evaluate the fit of factor models. Recently, we proposed the use dynamic fit index (DFI) cutoffs to evaluate model fit (McNeish & Wolf, 2021) and introduced a corresponding Shiny application to facilitate their use (Wolf & McNeish, 2020). Thus, a tutorial for its use in the applied research community is warranted. In this article, we will walk through 12 commonly asked questions about DFI cutoffs and use an applied example to demonstrate how to use the Shiny app to calculate them. For R users, DFI cutoffs are also available on CRAN under the package ‘dynamic’ (Wolf & McNeish, 2022).

1. What are the different types of model fit in Confirmatory Factor Analysis?

There are two types of *global* model fit in CFA: exact fit and approximate fit. Exact fit is a *test* of model fit in that it compares a test statistic to a probability distribution to calculate a *p*-value, while an approximate fit index can be thought of as an *effect size* measure that quantifies the degree of misfit in the model. Both are derived from the amount of overall misfit in the model, where misfit is defined as the difference between the model-implied variance-covariance matrix (e.g., the user’s path diagram) and the data-generated variance-covariance matrix (i.e., the observed relationships in the user’s data).

The most commonly used test of **exact fit** is the χ^2 test, although others can be used as well (see McNeish, 2020). Tests of exact fit are concerned with the presence of misfit (of any kind) anywhere in the variance-covariance matrices. The χ^2 test is a test of exact fit because the null hypothesis is that the model-implied variance-covariance matrix *exactly* matches the data-generated variance-covariance matrix. If the χ^2 test is significant, researchers have evidence to suggest that the model does not exactly fit the data. As such, the χ^2 test is the strictest way to evaluate model fit and test psychological theory (Hayduk et al., 2007).

The most used **approximate fit** indices are the standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), and the comparative fit index (CFI). There are no probability distributions or *p*-values associated with these indices, and thus they are not tests of fit but rather a way to evaluate the amount of misfit in the model (e.g., to determine if the misfit is trivial or substantial). In other words, approximate fit indices are different from tests of exact fit because they are concerned with quantifying the degree of misfit in the model. The SRMR is derived from the residual correlation matrix (i.e., the difference between the model-implied variance-covariance matrix and the data-generated variance-covariance matrix) and can be thought of as the average magnitude of the residuals. The RMSEA is derived from the χ^2 statistic but differs in that it includes a parsimony correction (i.e., it rewards simpler models with stronger theory that restrict more paths to 0). The CFI is the ratio of the model-reported χ^2 and the baseline χ^2 and can thus be thought of as the relative improvement in model fit. Lower values of SRMR and RMSEA are indicative of better fit while higher values of the CFI are indicative of better fit (Brown, 2015).

The last type of model fit is *localized* area of fit. In contrast to an overall evaluation of global fit, **local fit** is an investigation of each cell of the variance covariance matrix to diagnose any areas of strain in the model. This is often done to probe the source of misfit after finding a significant χ^2 statistic or a value of an approximate fit index that is indicative of substantial misfit. Still, good global fit can mask local misfit and thus it may be prudent to check regardless of global model fit. Although local fit is discussed at length in introductory textbooks for factor analysis (Brown, 2015; Kline, 2011), it is often not presented in journal articles.

Briefly, local fit can be investigated by probing the residual correlation matrix for extreme cases or by consulting modification indices for modifications that would substantially reduce the χ^2 statistic (Kline, 2011). Even if an extreme value is found, the model should not be changed unless the revisions are supported theoretically. There are several reasons for this. The first is that the modifications suggested by the software are derived statistically and may substantially alter the theory behind the initial model. Secondly, the misfit may be sample-specific and thus may not generalize to other samples. Third, there is no guarantee that a modification will resolve misfit and could instead lead a researcher down a rabbit hole akin to p-hacking. Fourth, it is difficult to know if misfit is due to an issue with the theory about the internal structure or one or more of the items. As such, any modifications should be justified theoretically and qualitatively.

2. What are the weaknesses of the current approaches to evaluate global model fit?

In terms of **exact fit**, the χ^2 test does not perform as well in small samples and may also be overly sensitive to minor misspecifications at large sample sizes (Browne & Cudeck, 1992; Hu et al., 1992). Additionally, with extremely small samples and smaller loadings, the χ^2 test can be underpowered and unable to detect misfit (McNeish et al., 2018). Some models (such as bifactor models) inherently have a higher “fit propensity”, meaning that they are more likely to fit the data regardless of the true nature of the data generating model

(Bonifay et al., 2017; Bonifay & Cai, 2017; Preacher, 2006). Further, even if one does find a non-significant p-value for a χ^2 test with a reasonable sample size, this does not guarantee that the true model has been recovered as there are often multiple equivalent models that can fit the data (Hayduk, 2014). As such, all modeling decisions should have strong theoretical grounding.

Unlike the χ^2 test, **approximate fit** indices do not have a corresponding *p*-value. Given that these indices essentially function as effect size measures that capture the degree of misfit in the model, the difficulty lies in how to interpret them and which cutoff values to use (if any). Researchers often rely on a set of fixed cutoff values derived from a simulation study conducted by Hu and Bentler (1999), which has over 96,000 citations as of 2022. This simulation study produced the well-known cutoff values of SRMR < .08, RMSEA < .06, and CFI > .95. However, interpretations of the results from simulations studies are limited to the conditions sampled in the simulation study. Hu and Bentler manipulated the sample size (from 250 – 5000), the number and type of misspecifications (omitted crossloadings and omitted factor covariances), and the normality of the factors and errors. However, they did not manipulate the number of factors (3), the number of items (15), the magnitude of the factor loadings (.7 - .8), or the model type (single level CFA estimated using maximum likelihood estimation).

Several studies have demonstrated that these fixed cutoff values cannot reliably be extrapolated to other model subspaces (e.g., one-factor models, multi-factor models with stronger or weaker loadings, or models with fewer or greater numbers of items or factors). In other words, if a researcher evaluates the fit of a single-level CFA model that does not have 15 items, 3 factors, and a sample size between 250 – 5000, *the cutoff values derived from Hu* *and Bentler's study cannot be used to reliably determine if there is substantial misfit in the model* (McNeish & Wolf, 2021). Most concerning is the "reliability paradox" which stipulates that lower loadings (e.g., a smaller reliability coefficient) are associated with "better" values of approximate fit indices (Hancock & Mueller, 2011; Heene et al., 2011; Marsh et al., 2004; McNeish et al., 2018; Saris et al., 2009). In other words, holding all else equal, as factor loadings decrease, the SRMR and the RMSEA will also decrease, mistakenly leading researchers to conclude that less reliable models fit the data better (when compared to a set of fixed cutoff values). If Hu and Bentler had varied the factor loadings in their original simulation study, it is possible that they would have been unable to recommend a set of fixed cutoff variables since fit indices are so sensitive to loading magnitude.

3. Why should I use DFI cutoffs instead?

Hu and Bentler's approach to quantifying the degree of misspecification in a model[6] was sensible; the problem lies with the interpretation of the fixed cutoff values (Millsap, 2007; Pornprasertmanit et al., 2013). The first problem, as mentioned above, is the extrapolation of this fixed set of cutoffs to conditions outside of the ones sampled in their original simulation. Secondly, because Hu and Bentler presented a single set of cutoffs, researchers were inadvertently encouraged to incorrectly treat misfit as a binary decision akin to a test of model fit (e.g., either the model fits well, or it does not). However, the only test of model fit is a test of exact fit; approximate fit indices are useful primarily because they can help researchers judge the extent to which the misfit in their model may be trivial or substantial.

[6] For an in-depth walk through of Hu and Bentler's derivation of misfit, see Hu and Bentler (1999) or McNeish and Wolf (2021).

DFI cutoffs are an improvement over the traditional fixed cutoff values because they address both of the issues raised above. DFI cutoffs are tailored to the user's specific model, which alleviates the first problem of improper extrapolation. Researchers can think of DFI cutoffs as "if Hu and Bentler had used my exact model for their simulation study, these are the cutoff values that they would have published". Unlike the traditional fixed cutoffs, DFI cutoffs (when available) are accurate for the user's model and can reliably distinguish between a correctly specified model and an incorrectly specified model. In this sense, DFI cutoffs can be thought of as analogous to a custom power analysis (albeit one that is quite simple to conduct).

Secondly, the DFI algorithm is written to return a series of custom cutoff values that range from trivial misfit to substantial misfit. This addresses the second problem of improper interpretation because it encourages researchers to treat misfit as a continuum or a spectrum rather than a binary decision of "good" or "bad". Because there is less finality associated with an interpretation of model fit when a series of cutoffs are used, the DFI approach also encourages researchers to properly reconceptualize model fit as only one type of validity evidence rather than the crux of validity (AERA, APA, NCME, 2014). As such, if a researcher found evidence of trivial misfit according to the DFI cutoffs, they could still potentially defend the use of their scale if they have other sources of evidence of validity in support of its use (while still acknowledging that if the χ^2 test is significant then the model does not exactly fit the data).

4. How does the DFI algorithm work?

The DFI algorithm follows Hu and Bentler's (1999) approach to model misspecification in that it simulates a distribution of fit indices from a correctly specified model and a misspecified model and then chooses a cutoff value that distinguishes between the two distributions. Hu and Bentler created their misspecified models by omitting one or two cross-loadings from a three-factor model (the “complex” condition; see Figure 1) and by omitting one or two factor correlations from the same three-factor model (the “simple” condition). The DFI algorithm mimics this approach by making the user’s model both the data generating model and the analytic model (this is the true condition). In the misspecified condition(s), the analytic model remains the user’s model, but a series of misspecifications are added to the data generating model (in line with the conventions established by Hu and Bentler). Readers interested in understanding the DFI algorithm in depth should consult McNeish and Wolf (2021; 2022).

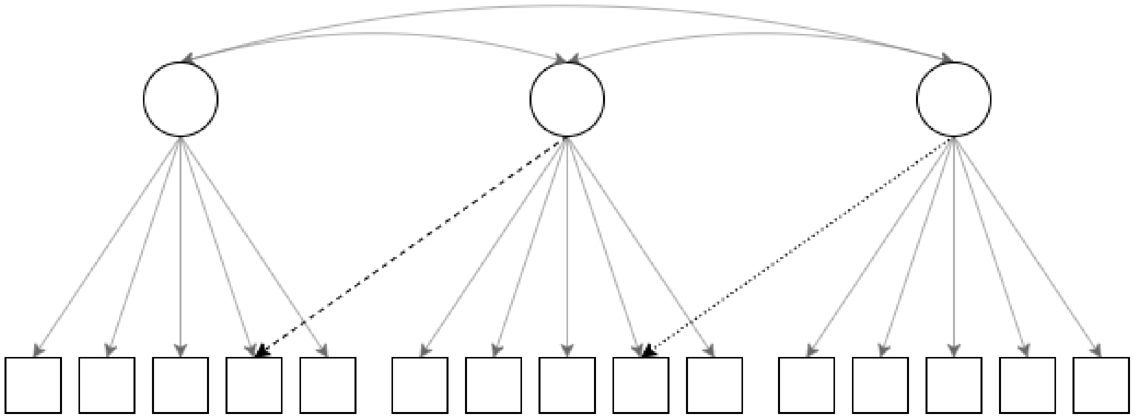


Figure 1. A path diagram of the model misspecifications used by Hu and Bentler (1999). The true data generating model has all of the loadings. The minor misspecification condition omits the loading with the dashed line from the data generating model, and the major misspecification condition omits both the dashed line and the dotted line from the data generating model. The analytic model is the same for all conditions.

Multi-Factor Models

For multi-factor models, the DFI algorithm creates *f*-1 levels of misspecifications, where *f* is the number of factors in the model. As such, a two-factor model will have one level of misspecification, while a six-factor model will have five levels of misspecification. The misspecification follows Hu and Bentler's approach of omitting a cross-loading with a magnitude equivalent to the lowest loading in the model from the factor with the highest reliability. For example, the Level 1 misspecification will omit one cross-loading with a magnitude equivalent to the lowest loading, while the Level 2 misspecification will omit both the Level 1 cross-loading and an additional cross-loading with a magnitude equivalent to the second lowest loading in the model. The magnitude of the loading that is omitted can be found in the Info tab of the app and will be returned by default in the R package.

One-Factor Models

The algorithm for one-factor models is unique in that it cannot exactly follow the approach established by Hu and Benter (1999) because it is impossible to omit factor correlations or cross-loadings in a one-factor model. As such, the DFI algorithm employs an approach inspired by Shi & Maydeu-Olivares (2020) of omitting residual correlations to create a misspecified model. One-factor models have three levels of misspecification[7], where Level 1 has approximately 1/3rd of items with an omitted residual correlation of .3, Level 2 has approximately 2/3rds of items with an omitted residual correlation of .3, and Level 3 has omitted residual correlations of .3 from all items. As such, the DFI algorithm for the one-factor model is standardized such that the number of items with an omitted residual correlation is proportional to the total number of items in the model, making it easier to compare degree of misfit across models.

[7] A four-item model will only have one level because there are not enough degrees of freedom to add additional levels. A five-item model will only have two levels because residual correlations are only able to be added to items that do not already have an existing residual correlation.

5. How do I calculate DFI cutoffs?

DFI cutoffs can easily be computed using the free, open source, web-based Shiny application, accessible at www.dynamicfit.app. The app has a simple, user-friendly, point-and-click interface, which requires no knowledge of coding to operate. The user need only enter their model statement with standardized loadings (see Question 6 for more details) and their sample size. Behind the scenes, the Shiny app will use R to run a series of Monte Carlo simulations to return a continuum of cutoff values tailored to the user's individual model. R users who wish to bypass the app can instead make use of the corresponding R package `dynamic`, available on CRAN, which will return the same results as the web application.

Applied Example

Computing DFI cutoffs is a post-hoc endeavor; in other words, users must first run their CFA model to get some of the information that is necessary to calculate custom fit index cutoffs. Thus, to make this tutorial easier to follow, we introduce an applied example which will be used throughout the rest of the paper. The data comes from a popular personality assessment commonly referred to as the "Big Five", which was provided by the Open Source Psychometrics Project (Goldberg, 1992). We will use the 10-item "extraversion" factor to compute DFI cutoffs for a one-factor model (see Figure 2).

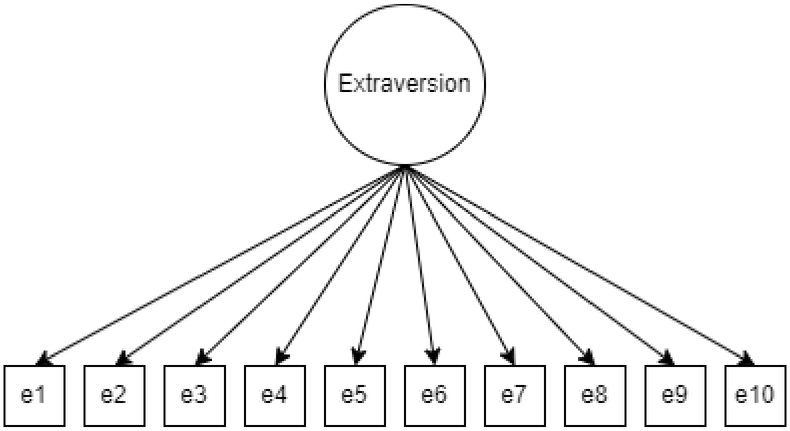
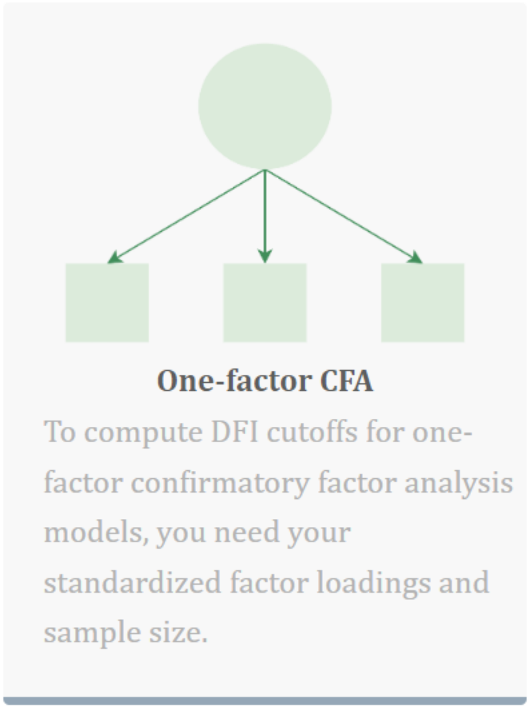


Figure 2. The one-factor model used for the demonstration in this tutorial (n = 1,222).

To use the Shiny application, researchers should visit the website and select the app that corresponds to their model type. In this case, we will select the one-factor CFA application (see Figure 3). The app description states that only two pieces of information are needed: (1) the user's standardized loadings from the fitted model, and (2) the sample size. The standardized loadings will be used to create the model statement which will be uploaded to the app to compute the custom DFI cutoffs (see Question 6). The function in the R package `dynamic` that corresponds to the one-factor CFA app is `cfaOne`.



One-factor CFA

To compute DFI cutoffs for one-factor confirmatory factor analysis models, you need your standardized factor loadings and sample size.

Figure 3. The one-factor CFA application on www.dynamicfit.app.

6. What does a model statement look like?

After opening the one-factor CFA app, users are prompted to enter two pieces of information: (1) their sample size, and (2) their model statement (see Figure 4). The model statement is created using the standardized loadings from the user’s fitted model (i.e., the results from the CFA model that the user wants to calculate DFI cutoffs for). These will be found in the software output that was used to run the original CFA model. We will walk through an example from Mplus and AMOS using the Extraversion factor from the Big 5 dataset provided.

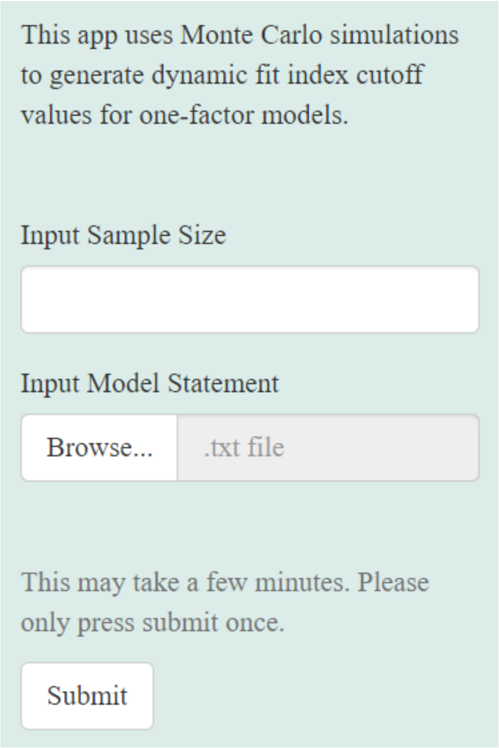
Mplus

To get the standardized loadings from Mplus (current as of version 8.7), we add the following argument to the end of the input file: `OUTPUT: STDYX;`[8]. The standardized loadings will be found under the section of the output titled `STDYX STANDARDIZATION` (see Figure 5). The magnitude of the standardized loading for each indicator is under the header titled `Estimate`. For example, the standardized loading for E1 is .671.

SPSS Amos

To get the standardized loadings from SPSS Amos (current as of version 28), users should select the “Analysis Properties” icon, and check the “Standardized estimates” box under the “Output” tab. After running the model, the standardized loadings can be found in the “Parameter Formats” box by deselecting “Unstandardized estimates” and selecting “Standardized estimates”. They will appear on the path diagram (see Figure 6a) and can easily be copied from the syntax. The syntax can be accessed by toggling to the “Syntax” tab underneath the path diagram (see Figure 6b).

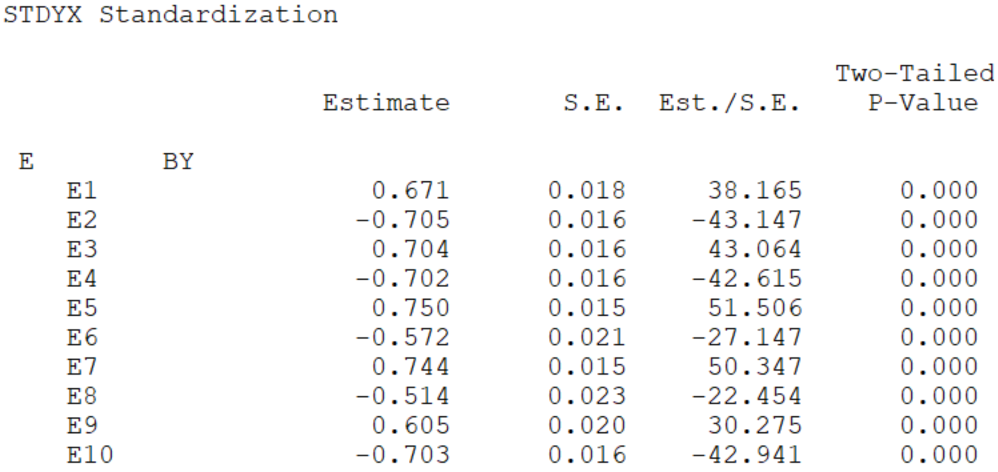
[8] We use STDYX because this standardizes both the latent variable(s) and the indicator variables (i.e., the items).



This app uses Monte Carlo simulations to generate dynamic fit index cutoff values for one-factor models.

This may take a few minutes. Please only press submit once.

Figure 4. The required inputs for the one-factor CFA app.

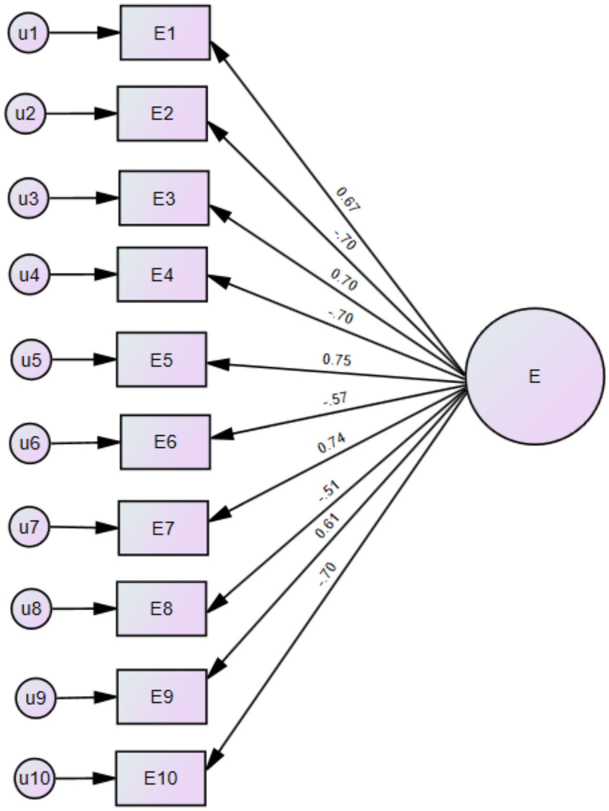


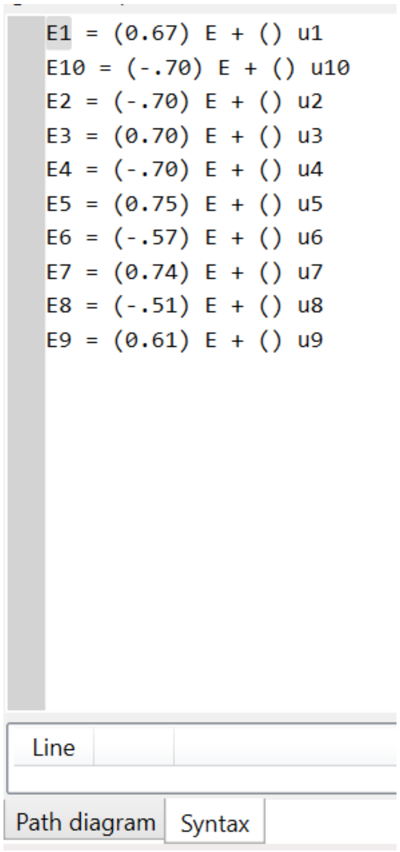
```
STDYX Standardization

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 E        BY
    E1                 0.671      0.018     38.165      0.000
    E2                -0.705      0.016    -43.147      0.000
    E3                 0.704      0.016     43.064      0.000
    E4                -0.702      0.016    -42.615      0.000
    E5                 0.750      0.015     51.506      0.000
    E6                -0.572      0.021    -27.147      0.000
    E7                 0.744      0.015     50.347      0.000
    E8                -0.514      0.023    -22.454      0.000
    E9                 0.605      0.020     30.275      0.000
    E10               -0.703      0.016    -42.941      0.000
```

Figure 5. The standardized loadings from Mplus.





```
E1 = (0.67) E + () u1
E10 = (-.70) E + () u10
E2 = (-.70) E + () u2
E3 = (0.70) E + () u3
E4 = (-.70) E + () u4
E5 = (0.75) E + () u5
E6 = (-.57) E + () u6
E7 = (0.74) E + () u7
E8 = (-.51) E + () u8
E9 = (0.61) E + () u9
```

Figure 6a. A path diagram with standardized loadings
loadings from Amos.

Figure 6b. The standardized

as seen in the Amos Syntax tab.

Model Statement

We will use the standardized loadings from Mplus and Amos to write out the model statement. Note that the model statement must be saved in a .txt file to be uploaded to the app. The easiest way to create a .txt file on a PC is in Notepad, while the easiest way on a Mac is in TextEdit (make sure to save your TextEdit file as plain text).

The model statement should be written in `lavaan` style syntax[9]. The regression relationship between the factor and any items will use the syntax `=~`, while any correlational relationships (e.g., between factors or items) will use the syntax `~~`. The model statement for this one-factor CFA model will be written as:

```
  Extraversion =~ .671*E1 + -.705*E2 + .704*E3 + -.702*E4 +
.750*E5 + -.572*E6 + .744*E7 + -.514*E8 + .605*E9 + -.703*E10
```

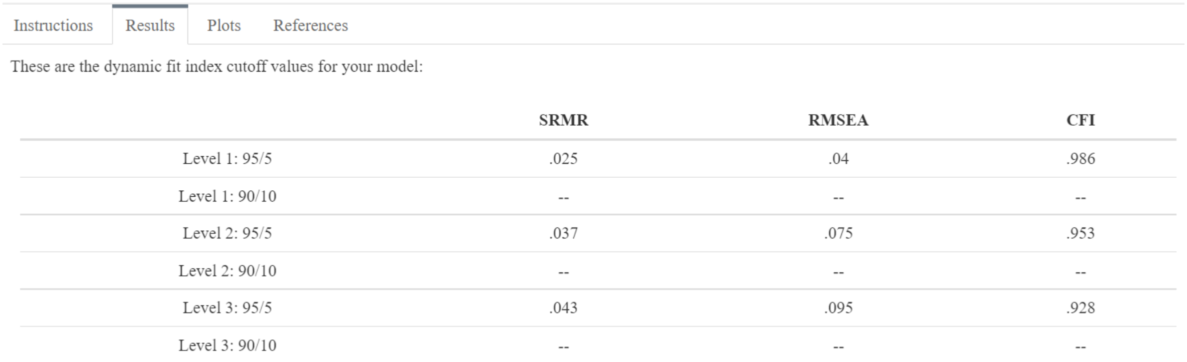
As seen above, the model statement follows the following format: Factor = item loading magnitude * item name. Because the first loading had a magnitude of .671, it is written as `.671*E1`. The magnitude of the loading will always come before the name of the item. The factors and items can have any name (e.g., `orange =~ .671*apple` would work), however the name cannot start with a number (e.g., `123orange` would not be permissible). Note that negative loadings still need to have a + sign as a link in the model statement (e.g., `.671*E1 + -.705*E2`). This model statement would then be saved as a .txt file and uploaded to the app along with the sample size (in this case, the sample size is 1,222). Users would then press submit to begin the simulation.

7. Why are there different levels and what do they mean?

After submitting the model statement and the sample size, the app will compute several Monte Carlo simulations (each with 500 replications) to return the DFI cutoffs. Once the busy bar finishes running, the DFI cutoffs will be found under the "Results" tab. For most models, there will be a series of cutoff values beginning with "Level 1". One-factor models

[9] `lavaan` is the latent variable modeling R package that runs behind the scenes in the Shiny app (Rosseel, 2012).

will typically have three levels of cutoff values[10]. Figure 7 displays the three levels of DFI cutoffs for the model used in this tutorial.



These are the dynamic fit index cutoff values for your model:

	SRMR	RMSEA	CFI
Level 1: 95/5	.025	.04	.986
Level 1: 90/10	--	--	--
Level 2: 95/5	.037	.075	.953
Level 2: 90/10	--	--	--
Level 3: 95/5	.043	.095	.928
Level 3: 90/10	--	--	--

Figure 7. The DFI cutoffs for the model used in this tutorial.

The levels correspond to increasing degrees of misspecification in the fit of the model, enabling researchers to reconceptualize misfit as a continuum analogous to an effect size measure. Misspecifications are cumulative such that higher levels are equivalent to more egregious model misspecifications. The Level 1 cutoff can therefore be thought of as the strictest fit criteria because it is consistent with the smallest misspecification from the misfit continuum. Thus, the ideal outcome would be if a user's fitted values for their SRMR, RMSEA, and CFI were all below the Level 1 cutoff, because this would indicate that the observed fitted values were more similar to a model that was correctly specified. In other words, if the user's model fit was below the Level 1 cutoffs, that would mean it fit better than a model that had only a minor misspecification. As the levels increase, the cutoff values will become more lenient because they will correspond to models that are more misspecified.

[10] Sometimes there will be fewer levels (e.g., when there not enough degrees of freedom to add an additional misspecification). For a detailed description of how the levels are derived, see McNeish and Wolf (2021).

If the fitted values for a one-factor model were above the Level 1 cutoff but below the Level 2 cutoff, one might conclude that their model was consistent with modest misspecifications and could potentially argue that the misspecifications were not a substantial threat to the validity of their inferences. If the fitted values were above the Level 2 cutoff but below the Level 3 cutoff, the model fit might be categorized as moderately misspecified, while fitted values above the Level 3 cutoff might be described as substantially misspecified. It is also possible to observe fitted values for each of the indices that are classified at differing cutoff levels. In this case, the researcher might report that the fit indices were consistent with different degrees of misspecification, and potentially attempt to diagnose the inconsistencies (e.g., by investigating local areas of strain). As always, it is up to the researcher to present multiple types of evidence in defense of the validity of their assessment (e.g., finding fit consistent with a minor misspecification should not be the entirety of the claim for evidence of validity). See Question 4 for more discussion about levels and model misspecification.

Applied Example

To determine how well their model fits their data, researchers should compare the fit of their model (derived from their software of choice) to the DFI cutoffs derived from the app or the R package. In this case, the fit index values for the Extraversion empirical model are $\chi^2 = 436.55$ (df = 35, p < .001), SRMR = .048, RMSEA = .097 [.089, .105], CFI = .921. Compared to the traditional fixed cutoff values from Hu and Bentler (1999), the SRMR would be indicative of good fit while the RMSEA and the CFI would be indicative of poor fit.

There are two drawbacks to using the fixed cutoffs. The first is that the fixed cutoffs are derived from a different model subspace (a 3-factor model with 15 items) that does not generalize to our model (a one-factor model with 10 items). The second is that because there is only one set of cutoff values, we cannot infer the degree of misspecification present, forcing us to make a binary decision about a continuum of misspecification. Alternatively, when comparing our fit index values to the DFI cutoffs that are tailored to our empirical model, we see that all three indices are consistent with a Level 3 misspecification (see Figure 7). This is because the SRMR and RMSEA are greater than the Level 3 DFI cutoff, while the CFI is less than the Level 3 DFI cutoff. As such, we can infer that the model is substantially misspecified, rendering it difficult to defend as valid especially without other types of validity evidence. We might follow up by investigating local fit such as consulting the modification indices in our software of choice. In doing so, we see that adding a residual correlation between E8 and E9 would reduce the χ^2 by 156.158, or 36%. The next steps might involve qualitatively investigating the cause of the relationship between these two items, removing one of them if they are deemed redundant, adding a residual correlation between the two of them if it is theoretically justified, or something else.

8. What does 95/5 and 90/10 mean, and how should I interpret the plots?

The 95/5 and 9/10 thresholds are derived from Hu and Bentler's approach to minimizing the classification error rates. In Question 4, we mentioned that the DFI algorithm simulates a distribution of fit index values from a correctly specified model and a misspecified model and then selects a cutoff value that distinguishes between the two distributions. In the app, these distributions are visualized under the "Plots" tab (see Figure 8 for the Level 3 distributions from the applied example). But how is that value selected, especially when the distributions might overlap leaving researchers unsure as to whether the fit index value that they observed is consistent with one that is derived from a misspecified model or a correctly specified model?

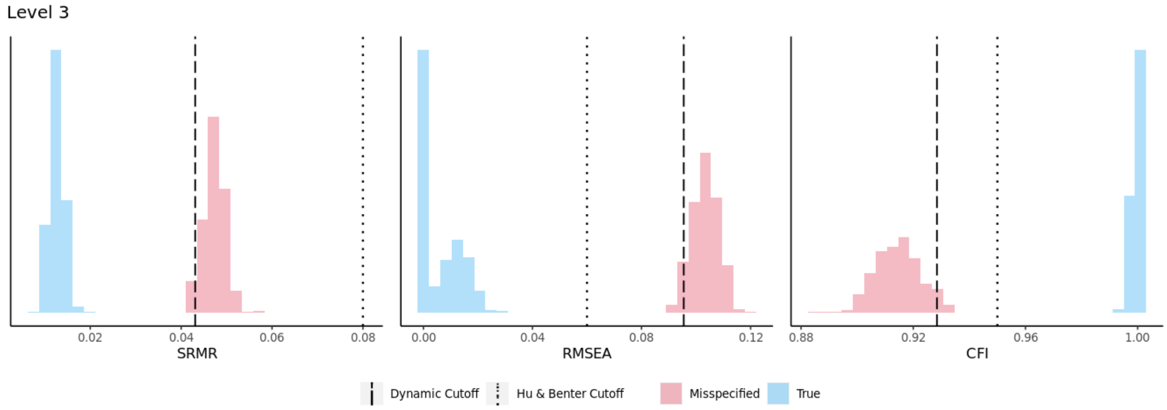


Figure 8. The Level 1 distributions from the DFI algorithm for the Extraversion scale.

To avoid ambiguity, the DFI algorithm consistently selects the cutoff value from the misspecified distribution. Specifically, the magnitude of the cutoff value corresponds to the 5th percentile of the misspecified distribution for the SRMR and RMSEA and the 95th percentile of the CFI (this is because low values of SRMR and RMSEA are indicative of better fit while high values of CFI are indicative of better fit). These values correspond to the dashed lines in Figure 8 (the Hu and Bentler cutoffs are also presented for comparison as dotted lines). Since the fitted value for the SRMR in the Extraversion example was .048, it is more likely that this value would come from a distribution of misspecified fit indices because the SRMR misspecified distribution (in red) ranges from .041 to .056, while the SRMR correctly specified distribution (in blue) ranges from .008 to .019. As such, the conclusion is that the fitted model is likely misspecified in a way that is consistent with a Level 1 misspecification.

In Figure 8, the two distributions are distinct and clearly separated. However, sometimes the misspecified and the correctly specified distributions will overlap. Since the distributions can overlap, the rule that the DFI algorithm uses is to select the 5th percentile of the misspecified distribution (for the SRMR and RMSEA) *so long as* the value that is returned is also greater than the 95th percentile from the correctly specified distribution[11]. This check is put in place to safeguard against mistakenly choosing a cutoff value that could just as easily come from a correctly specified distribution. If the distributions overlap too much, (such that the 5th percentile of the misspecified distribution is less than the 95th percentile of the correctly specified distribution), then the DFI algorithm will attempt to return the 10th percentile of the misspecified distribution so long as the value is greater than the 90th percentile of the correctly specified distribution. Conceptually, this is like changing the alpha from .05 to .10 in standard null hypothesis significance testing. As such, the probability of misclassification is higher with the 90/10 rule than with the 95/5 rule, but the overall likelihood of making an error is still reasonably low (McNeish & Wolf, 2022).

9. What does NONE mean, and what should I do if I see it?

In Question 8, we spoke about how the cutoff values are derived from the 5th percentile of the misspecified distributions of fit indices (for SRMR and RMSEA; the 95th percentile for CFI), and the problems that begin to arise if the misspecified and correctly specified distributions overlap. If the misspecified and correctly specified distributions of fit indices overlap substantially, then we become unsure as to whether an observed fitted value would be more likely to be found in a distribution of fit indices that were derived from a

[11] The opposite is true for the CFI (e.g., the DFI algorithm will select the 95th percentile of the misspecified distribution so long as the value that is returned is smaller than the 5th percentile of the correctly specified distribution).

correctly specified model or a misspecified model. An example of this can be seen in Figure 9. When this happens, the DFI algorithm will return the word “NONE” for that fit index.

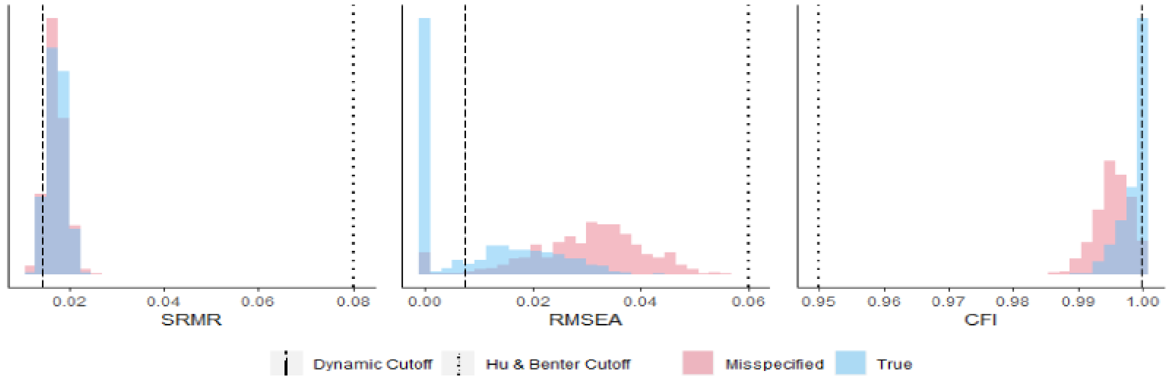


Figure 9. An example of overlapping distributions that would result in a NONE outcome.

When the word “NONE” is returned, that means that there are no cutoff values for that level of misfit that can reliably distinguish between a correctly specified model and a misspecified model. This can be verified visually in the “Plots” tab of the app. This is more likely to happen when sample sizes are small and loadings are low (Hancock & Mueller, 2011; Heene et al., 2011; McNeish et al., 2018). If there are DFI cutoff values available for other indices or other levels, they can and should still be used. If there are no DFI cutoff values available for any indices or any levels, the solution is not to rely on the traditional fixed cutoff values from Hu and Bentler as they similarly cannot distinguish between a correct and misspecified model. Instead, users can attempt to collect more data to increase their sample size, rely on the χ^2 test, or investigate local fit.

10. How do I include DFI cutoffs in a manuscript?

In our experience, it is easiest to include DFI cutoffs in a manuscript by putting them in a table and then referencing the table in the text of the article. Researchers should report the fit of their model as they normally would, and then reference the likely magnitude of misspecification by comparing each approximate fit index to the table of DFI cutoff values. An example of a write up using the model from this tutorial is presented in the next section. Because tailored cutoffs (specifically, DFI cutoffs) are a relatively new development, it may be worthwhile to mention that they are used to quantify the degree of misfit in the model or include a sentence about the limitations of fixed cutoff values for readers who are not familiar with existing literature.

In this tutorial, the model was substantially misspecified. When this happens, researchers may be interested in modifying the model to attempt to improve the fit (although, note that model modification is only potential resolution and modifications should not be made without strong theoretical justifications). It is not clear to us quite yet how to proceed with DFI cutoffs when using modifying the model (i.e., we are not sure if DFI cutoffs should be updated for the new model or not). This is discussed more in Question 12. We are currently working on resolving this conundrum so that we can make clear recommendations for researchers.

Applied Example: Standard Reporting

The test of exact fit was statistically significant (χ^2 = 436.55, df = 35, p < .001) indicating that the model did not exactly fit the data. The approximate fit indices for the model were SRMR = .048, RMSEA = .097 [90% CI (.089, .105)], and CFI = .921. These indices are essentially effect size measure for the magnitude of misfit. To quantify the degree of misfit reflected in these indices is, we compare the fit indices to a series of dynamic fit index (DFI) cutoffs (McNeish & Wolf, 2021) calculated by the one-factor DFI Shiny app version 1.1.0 (Wolf & McNeish, 2020). A table with the resulting cutoffs derived for this model are shown below. The SRMR and RMSEA from the model were above the Level-3 DFI cutoff and the CFI was below the Level-3 DFI cutoff, indicating that the fit of the model is consistent with a substantial misspecification.

Table A. The DFI cutoffs used to quantify the degree of misfit in the model.

	SRMR	RMSEA	CFI
Level-1	.025	.040	.986
Level-2	.037	.075	.953
Level-3	.043	.095	.928

11. What should I do if DFI cutoffs don’t exist for my model type?

As of this writing, the DFI method supports CFA models with continuous indicators only and the simulation component of the software assumes multivariate normality. It takes some time to work out generalizations for other types of models because a general method for identifying relevant misspecifications must be done model by model. For instance, potential misspecification that are relevant to latent growth models would likely be very different than a confirmatory factor analysis because latent growth models are typically interested in aspects like the function form of growth being correct or whether the correlation between repeated measures is reasonable rather than things like omitted cross-loadings that are relevant to confirmatory factor analysis. Our current work is focused on extending the method to higher-order models, categorical indicators, non-normality, missing data, and measurement invariance; so we expect those or related extensions will be the next to be added to the DFI method.

In the meantime, researchers should rely on the chi-square test and investigate local areas of strain to look for obvious misfit (e.g., viewing the standardized residual covariance matrix). Researchers can also use the Exact Fit application in the Shiny App or the `exactFit` function in the R package to return the 95th or 99th percentile of the distribution of cutoff values for a correctly specified model (for researchers that used one of the currently available apps, this can also be found in the Level 0 tab). Because this is a distribution of cutoff values for the true model, the values that are returned are the strictest way to evaluate approximate model fit. If the researcher's fit index values fall below these values, this indicates that the fit of their model is consistent with a model that is correctly specified. These cutoff values can be computed for any model with continuous outcomes (e.g., models estimated using ML or MLR).

12. What are the limitations of DFI cutoffs?

In addition to only being available for a limited number of model types, DFI cutoffs have three other notable limitations. Currently, the cutoff values are derived by simulating data that is multivariate normal, which may not be consistent with the researcher's real data. We are currently working on switching to a bootstrapping approach which would sample the researcher's data and account for any non-normality anywhere in the model, resulting in cutoff values that are more accurate. Implementing this would require researchers to upload their data to the app, but it would mean that the model statement was simpler to write (e.g., it would no longer be necessary to include the magnitude of the standardized loadings in the model statement).

Additionally, the misspecifications for the one-factor model are currently standardized and thus somewhat comparable across models regardless of the number of items, but the misspecifications for the multi-factor models are not. This is because the multi-factor model replicates Hu and Bentler's approach to misspecification which involves adding one cross-loading for each *f*-1 factor in the model with a magnitude equivalent to the item with the lowest loading in the model. Meanwhile, the one-factor model simply adds a series of residual correlations with magnitudes of .3 proportional to the total number of items in the model. We are working on introducing a similar standardized approach to model misspecification for multi-factor models.

Lastly, it is not clear to us quite yet if DFI cutoffs should be recomputed every time a model is modified. It is possible that cutoff values could change considerably for multi-factor models if the magnitude of the lowest factor-loading changes substantially (e.g., from .3 to .7). This will likely be resolved by standardizing the approach to misspecification for multi-factor models, which will make it easier to determine if the cutoff values should be recalculated. At this point, we hypothesize that it may not be necessary to recompute DFI cutoffs for small modifications to a model (i.e., adding a residual correlation) but it may be necessary to recompute DFI cutoffs for larger modifications (e.g., switching from a one-factor model to a two-factor model).

References

AERA, APA, NCME. (2014). Validity. In *Standards for Educational and Psychological Testing* (pp. 11–31). American Educational Research Association.

Bonifay, W., & Cai, L. (2017). On the Complexity of Item Response Theory Models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science, 5*(1), 184–186. https://doi.org/10.1177/2167702616657069

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). The Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2). https://journals.sagepub.com/doi/10.1177/0049124192021002005

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Hancock, G. R., & Mueller, R. O. (2011). The Reliability Paradox in Assessing Structural Relations Within Covariance Structure Models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Hayduk, L. (2014). Seeing Perfectly Fitting Factor Models That Are Causally Misspecified: Understanding That Close-Fitting Models Can Be Worse. *Educational and Psychological Measurement*, *74*(6), 905–926. https://doi.org/10.1177/0013164414527449

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! Testing! One, two, three – Testing the theory in structural equation models! *Personality and Individual Differences*, *42*(5), 841–850. https://doi.org/10.1016/j.paid.2006.10.001

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362. https://doi.org/10.1037/0033-2909.112.2.351

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed.). The Guilford Press.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

McNeish, D. (2020). Should We Use F-Tests for Model Fit Instead of Chi-Square in Overidentified Structural Equation Models? *Organizational Research Methods*, *23*(3), 487–510. https://doi.org/10.1177/1094428118809495

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*.

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

Pornprasertmanit, S., Wu, W., & Little, T. D. (2013). A Monte Carlo Approach for Nested Model Comparisons in Structural Equation Modeling. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New Developments in Quantitative Psychology* (pp. 187–197). Springer. https://doi.org/10.1007/978-1-4614-9348-8_12

Preacher, K. J. (2006). Quantifying Parsimony in Structural Equation Modeling. *Multivariate Behavioral Research*, *41*(3), 227–259. https://doi.org/10.1207/s15327906mbr4103_1

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Shi, D., & Maydeu-Olivares, A. (2020). The Effect of Estimation Methods on SEM Fit Indices. *Educational and Psychological Measurement*, *80*(3), 421–445. https://doi.org/10.1177/0013164419885164

Wolf, M. G., & McNeish, D. (2020). *Dynamic Model Fit* (1.1.0) [R Shiny]. https://www.dynamicfit.app

Wolf, M. G., & McNeish, D. (2022). *dynamic: DFI cutoffs for latent variable models* (1.1.0) [R]. https://cran.r-project.org/web/packages/dynamic/index.html

Paper 3: The Problem with Over-Relying on Quantitative Evidence of Validity

Introduction

Self-report survey instruments are ubiquitous throughout the social sciences (Maul, 2017). They are commonly found in social and personality psychology (e.g., the Big Five; McCrae & Costa, 2008), clinical psychology (e.g., the depression scales such as the CES-D; Radloff, 1977), medical research (e.g., patient reported outcomes such as the PROMIS; Cella et al., 2010), educational research (e.g., self-control and grit; Duckworth et al., 2007), and in public opinion polling, among others. These scales can be used in high stakes settings such as measuring treatment efficacy in medical trials, evaluating the suitability of applicants for military or corporate careers, determining the amount of funding allocated by state or federal budgets, or simply to contribute to the current state of knowledge of psychological theory and practice through research studies. Self-report survey instruments are an important vehicle for generating data about humans and making decisions that impact individuals or groups of individuals.

Self-report surveys are prima facie simple to create and disseminate. Anyone with access to the internet can write a question, post it on a social media website as a poll, collect data, and use the results to make decisions. However, simply creating a survey item or set of survey items and collecting responses does not mean that the survey author can draw meaningful conclusions or make accurate, evidence-based decisions. Professional assessments can be distinguished from fun, amateur online personality quizzes through subjugation to a series of rigorous tests and quality checks in a process known as validation.

The practice of validation can vary across academic disciplines, but the general goal is to present evidence that demonstrates the adequacy and appropriateness of the use of a scale for a particular purpose (Maul, 2018; Messick, 1989; Zumbo & Chan, 2014)[12].

The type of validity evidence that is presented depends on the kind of survey that is created and the intended interpretation of the survey results (Kane, 2013). It is common for social scientists to create surveys to measure unobservable psychological attributes such as anxiety, conscientiousness, or self-efficacy. To do this, survey authors construct multiple items that address all relevant aspects of the psychological attribute and collect responses from a sample of participants that are representative of the population the author intends to study. If the responses to the items are indeed caused by same psychological attribute, the survey author would expect to see a similar response pattern across the items. This is commonly tested using a latent variable model known as confirmatory factor analysis (CFA).

A CFA model functions as a mathematical representation of a theorized causal process where it is hypothesized that the responses to survey items are caused by a psychological attribute. Within a CFA model, the *factor* is the latent or unobservable construct that a researcher intends to measure (e.g., growth mindset), and the survey *item* responses are the manifestations of that psychological attribute (e.g., responses to survey items about growth mindset). In a one-factor model, the dependent variable is the item response, and the independent variables are the strength of the relationship between the item and the factor (the *factor loading*), the latent variable value, and the item mean. Researchers can then test if this theorized causal model is plausible by consulting several indices of *model*

[12] For a noteworthy dissent, see Borsboom, et al. (2004).

fit (e.g., χ^2, SRMR, CFI) to determine if the CFA model adequately replicates the observed item responses (Brown, 2015). If the statistical measurement model adequately fits the observed data, researchers may claim that the scale is valid because the survey items conform to the construct on which the proposed survey uses are based (AERA, APA, NCME, 2014; Tarka, 2018).

CFA models have become a prominent source of validity evidence in psychology, appearing in up to 85% of published scale validation journal articles (Chinni & Hubley, 2014; Cizek et al., 2008; Hubley et al., 2014; Slaney, 2017). Given their importance in the practice of survey validation, one might assume that evaluating the fit of a CFA model is a settled issue as it is often the decisive element in evaluating if a psychological attribute has been successfully measured by a survey instrument. However, unbeknownst to many applied social science researchers, there is widespread disagreement among quantitative methodologists and psychometricians over how to define whether the fit of a CFA model to observed data is adequate (McNeish et al., 2018).

CFA model fit is often evaluated using approximate fit indices (AFI) such as SRMR, RMSEA, and CFI. AFI's are often used because tests of exact fit (e.g., χ^2) are significant, indicating that the model does not exactly fit the data (Hayduk, 2014). AFI's essentially function as effect size measures in which researchers attempt to quantify the degree of misfit in the model (McNeish et al., 2018; Millsap, 2007). Unlike tests of exact fit, AFI's are not associated with any distributional assumptions which makes it difficult to interpret their magnitude (Mulaik, 2007; Saris et al., 2009). Most often, researchers rely on a set of fixed cutoff values from Hu and Bentler's (1999) simulation study to interpret them.

Part of the disagreements over CFA model fit stems from the improper application of AFI's as a *test* of model fit and the lack of recognition of the distinction between perfect fit and degree for misfit. The second (and perhaps more notorious) is that the cutoffs established by Hu & Bentler (1999) are not only treated as a test of fit, but also improperly generalized across all model subspaces and do not properly distinguish between correctly and incorrectly specified models outside the condition(s) in which they were derived (Heene et al., 2011; Marsh et al., 2004; McNeish et al., 2018). The commonly used fixed cutoffs values originating from Hu & Bentler's (1999) seminal work were derived from simulations of one single-level CFA model: a three-factor model with 15 items, all with loadings ranging from .70 to .80. They were designed to detect the absence of one missing cross-loading with a magnitude equivalent to the value of the item with the lowest factor loading. The authors clearly stated that the cutoffs derived from their simulation work should not be applied across all model subspaces, but their caution was not heeded, as their paper has more than 95,000 citations to date, making it one of the most popular papers in all of psychology.

Recently, McNeish and Wolf (2021, 2022) introduced Dynamic Fit Index (DFI) cutoffs; a simulation-based approach to generating fit index cutoffs that are tailored to the user's specific factor model. They created an algorithm for multi-factor models that replicates the misspecification method used by Hu and Bentler (1999) to identify cutoff values that reliably distinguish between misspecified and correctly specified models. This algorithm is easily accessible via a shiny app (www.dynamicfit.app) or it can be found in the R package `dynamic`. The DFI algorithm allows researchers to pretend their model was Hu and Bentler's data generating model in their simulation study, and essentially returns the cutoff values that Hu and Bentler would have published if they had used the researcher's model in their study instead. Given the frequency in which CFA models are used as evidence of validity and the fact that model fit is typically a defining factor in claims of validation, it seems warranted to re-evaluate the strength of these validity claims using cutoff values that are appropriate for the model under consideration.

The Present Study

There are thousands of survey instruments in the social sciences to choose from for this review; more than 280 measures of depression severity have been created alone in the last 100 years (Santor et al., 2006). We chose to focus on some of the most well-known and influential survey instruments: those with at least 5,000 citations for the original validation article. Using this inclusion criteria means that older scales are more likely to be eligible for inclusion than recently published scales as enough time has passed for them to collect 5,000 citations. However, older scales were also less likely to use CFA as evidence of validity simply because the computing power and software necessary to run CFA models was not yet commercially available and prior to Hu and Benter (1999) there was no common consensus on how to interpret AFI's (Marsh et al., 2004; Slaney, 2017). Instead, researchers often relied on quantitative metrics such as reliability coefficients (e.g., coefficient alpha), consistency across multiple administrations of a scale (e.g., test-retest correlations), and relationships with other variables (e.g., criterion validity). For example, a meta-analysis of the Big Five found 231 criterion-validity studies conducted between 1952 and 1988 (Barrick & Mount, 1991).

Thus, if the original validation article for a survey instrument did not include a CFA as evidence of validity, we included two journal articles about it in this review: the original one, and a more recent one that used a CFA model as evidence of validity. Because DFI cutoffs were created to replicate Hu and Bentler's approach to model misspecification, only single-level CFA models with continuous outcomes (i.e., estimated using ML or MLR) are eligible for inclusion in this study. Additionally, two pieces of information are needed to calculate model specific DFI cutoffs: the factor loadings, factor correlations, and residual correlations for all items and factors in the model, and the sample size. This information was surprisingly difficult to find. As such, scales or publications that used CFA but did not provide this information were excluded. In addition to recalculating model specific DFI cutoffs and retesting model fit, we also report on the types of validity evidence that were presented, and whether the definition of the psychological attribute, the intended use of the scale[13], and scoring instructions were clearly articulated by the survey developers.

Results

We investigated nine well-known scales in this review, each with more than 5,000 citations (see Table 1). The oldest scale was constructed in 1960 and the most recent scale was created in 2007 (median = 1988). The median number of citations was 17,081 (min = 5,934, max = 48,618). On average, 41.2% of these citations came from papers written in the last four years. The number of citations was gathered from Google Scholar.

Table 1. The scales covered in this review, ordered by year of publication.

Scale	Abbreviation	First Published	Number of Citations (2022)	Number of Citations since 2018
Beck Depression Inventory (Beck et al., 1961)	BDI	1961	48,618	9,390
Maslach Burnout Inventory (Maslach & Jackson, 1981)	MBI	1981	20,363	9,080
Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983)	HAD	1983	46,828	15,800

[13] Uses other than "measurement for the sake of measurement"

Edinburgh Postnatal Depression Scale (Cox et al., 1987)	EPDS	1987	13,799	5,880
Positive and Negative Affect Schedule (Watson et al., 1988)	PANAS	1988	47,945	16,100
The Pittsburgh Sleep Quality Index (Buysse et al., 1989)	PSQI	1989	28,721	14,500
Pain Catastrophizing Scale (Sullivan et al., 1995)	PCS	1995	7,288	3,490
Ten-Item Personality Inventory (Gosling et al., 2003)	TIPI	2003	9,389	4,750
Grit Scale (Duckworth et al., 2007)	Grit-O	2007	7,763	5,020

Quantitative evidence of validity was presented more often than qualitative evidence (see Table 2). Reliability coefficients were presented for every scale, as was evidence of correlational relationships with external criterion (commonly described as predictive or discriminant validity). Slightly more than half presented some evidence based on the internal structure, typically in the form of a principal components analysis (PCA). Most scales gave a brief one-sentence description of how they selected the items for their scales (e.g., a literature review) which falls under content evidence (none of the authors wrote more than two or three sentences about this). Two scales briefly referenced response process evidence (e.g., pretesting items for interpretability) and none of the scales discussed the consequences resulting from the use of the scale. Slightly more than half offered a clear definition of the properties the scale was designed to measure, while slightly less than half stated the intended uses of the scale or gave instructions on how to score it.

Table 2. Types of validity evidence presented in the introductory article for each scale.

	Quantitative Evidence			Qualitative Evidence					
Scale	Reliability	Criterion	Internal Structure	Content	Response Process	Consequences	Attribute Definition	Intended Use	Scoring Instructions
BDI	X	X		*				X	
MBI	X	X	X	*			X		
HAD	X	X							X
EPDS	X	X	X	*	*			X	X
PANAS	X	X	X	*			X		
PSQI	X	X		*			X	+	X
PCS	X	X	X	*	*		X	X	X
TIPI	X	X							
Grit-O	X	X	X	*			X		

+ The PSQI stated that it was created to identify good/bad sleepers, which we characterized as measurement for the sake of measurement. However, they subsequently described how the scale could potentially be used in a clinical setting.
* This was addressed briefly in one to three sentences.

Only one scale (Grit-O) used CFA as evidence of validity, but the authors did not provide the factor loadings necessary to compute DFI cutoffs[14]. Thus, we used the results of CFA models published in more recent validation studies for all scales (all of which were published in 2000 or later). The fact that we were able to easily find studies that presented such evidence of validity speaks to the frequency in which CFA analyses are currently relied on as evidence of validity. All studies that reported chi-square tests reported significant chi-square values, indicating that the hypothesized model did not exactly fit the data. All of the studies published at least two AFI values (RMSEA and CFI were most common). The AFI values for all of the studies met at least one of the AFI cutoff values established by Hu and Bentler[15] (usually two), although nearly all of the authors modified the survey instruments from their original form to achieve adequate model fit (see Table 3). Most changed the dimensionality of the scale, while others removed items, added cross-loadings or error correlations, or both. Only one of the scales met the revised DFI cutoff values (after substantial modification from the original published version). However, DFI cutoffs were not available for four of the scales in this study (noted with an asterisk in Table 3). This commonly occurs when factor loadings are too low or the sample size is too small (this can be thought of as the model not being sufficiently powered to distinguish between correctly specified and misspecified models)[16].

[14] A revised version of the Grit Scale, the Grit-S, was released shortly after the first version was published. Thus, we used the Grit-S to compute DFI cutoffs. While the revised version did include the information necessary to compute DFI cutoffs, it was unusable because one of the standardized loadings exceeded 1.0, which is an impossible value.

[15] Hu and Bentler recommended a CFI cutoff value of .95. However, researchers often rely on a more lenient cutoff value of .9 from one of Bentler's earlier publications (Bentler & Bonett, 1980).

[16] If Hu and Bentler had used a data generating model with lower loadings in their original study, they likely would not have been able to make any cutoff recommendations.

Table 3. The fit of each model compared to Hu and Bentler’s cutoffs and DFI cutoffs.

Scale	Same structure	Changed Dimensionality	HB: SRMR	HB: RMSEA	HB: CFI	CFI (> .9)	DFI: SRMR	DFI: RMSEA	DFI: CFI
BDI (Whisman et al., 2000)			--	X		X	--	*	*
MBI (Vanheule et al., 2007)			X	X		X		*	*
HAD (Dunbar et al., 2000)		X	--	X	X	X	--	X	X
EPDS (Coates et al., 2017)		X	--	X	X	X	--		
PANAS (Galinha et al., 2013)			X	X		X	X		
PSQI (Gelaye et al., 2014)		X	--	X	X	X	--		
PCS (Sehn et al., 2012)	+	X	--		X	X	--	*	*
TIPI (Muck et al., 2007)	X		--		X	X	--	*	*
Grit-S (Gonzalez et al., 2020)	+	X	X		X	X			

-- A value for this fit index was not reported in the journal article

+ This structure was only modified slightly

* DFI cutoff values were not returned for this model (likely because the loadings were too low or the sample size was too small)

Discussion

Most of the scales in this review were published in the late 1900s, when psychology was primarily concerned with evaluating instrument quality by calculating reliability coefficients and estimating the relationship between their survey instrument and scales designed to measure other relevant properties. This was partially due to the popularity of the nomological network (Cronbach & Meehl, 1955) and the conceptualization of validity at the time (discussed more in the next section). Thus, it is not surprising to see that most of the evidence presented was quantitative in nature. It is possible that qualitative evidence of validity has since been presented in other studies; a meta-analysis would need to be conducted for each scale to resolve this. However, Chinni and Hubley (2014) conducted a meta-analysis of the popular Satisfaction with Life Scale (SWLS; Diener et al., 1985) and found that out of 46 studies, only one presented any qualitative evidence of validity while the rest presented at least one type of quantitative evidence.

Approximately half of the scales in this study included a definition of the attribute, a scoring rubric for the scale, or intended uses of the scale. The described uses for all scales were somewhat vague (e.g., “clinical or diagnostic purposes”) and scoring instructions (when included) were typically in the form of a total sum score across multiple factors. This does not mean that scoring instructions for some scales were not included in subsequent publications (e.g., the BDI-II included scoring instructions in its manual). Nonetheless, ambiguity in score meaning or appropriate scale use and uncertainty in proper scoring makes it more likely that the scales may be misused or interpreted inconsistently. It is worth noting that on average, 40% of the citations for these studies were from articles published in the last 4 years indicating that these scales likely continue to be used in the literature, if not just as external criterion to validate new scales. If the validity of these scales is in question, it may result in biased conclusions about the validity of scales that rely on correlations with these scales to substantiate their use.

All of the scales in this study did not present detailed evidence that participants understood items as intended or engaged in the expected cognitive processes when responding to survey items (e.g., evidence based on the response process; Castillo-Díaz & Padilla, 2013; Zumbo & Hubley, 2017), although two scales briefly referenced pretesting items before administering them. Further, none of the scales published a detailed analysis of why these items were chosen instead of other items or how they mapped onto the property of interest. This begs the question: why use *these* items? Would we still use these scales if they had been introduced today? Would they meet the standards for today's best practices of validation? Given that quantitative best practices may evolve over time, strong measurement theory and qualitative evidence that participants understand items as intended may help researchers defend the continued use of their scales.

All of the scales published in this review (that reported chi-square values) had statistically significant chi-square values, indicating that the hypothesized measurement models did not exactly fit the data. However, the AFI values for all the scales met at least one (usually two) of the criteria established by Hu and Bentler, signifying that these improperly generalized cutoffs values may serve a key deciding factor in establishing evidence of validity in psychology. Only two of the scales met the DFI cutoffs (i.e., revised Hu and Bentler cutoffs that were tailored to the user's model). Thus, any scales that rely exclusively on CFA as evidence of validity and modified their models to meet Hu and Bentler's cutoff criteria may find that the validity of their scale is now in question. Nearly all of the CFA models in this study deviated from the hypothesized structure in the original validation articles, ostensibly to find a better fitting model. Thus, it seems that psychologists may consider the structure of the construct to be more malleable than the items themselves and that the goal may be to modify the model until desirable psychometric properties are found. This approach seems somewhat unusual given that none of the scales presented detailed evidence of individual item quality (e.g., evidence that participants understood items as intended). When a factor model does not fit the data, it can be because either the theory is not sound or the items require revision; the cause cannot be determined by statistics alone. Electing to modify the psychometric model instead of investigating item quality and the cause of misfit is akin to allowing statistics to guide scale construction instead of allowing theory to guide psychological measurement.

This is not to say that the validity of the scales that were reanalyzed in this review has been refuted. Validation is an on-going process for which multiple sources of evidence can be presented depending upon the validity argument one wishes to make (AERA, APA, NCME, 2014; Messick, 1990). The social science literature is quite vast, and it is possible that these samples were outliers and adequate fit may be found in a different sample. Further, it was only possible to compute DFI cutoffs for these scales because the authors were so thorough in their reporting, which is commendable. Many scales (such as the Big Five; McCrae & Costa, 2008) were not able to be included in this study because they did not present the information required to calculate DFI cutoff values. Additionally, items for many scales (e.g., the MMPI; Graham, 1987) are treated as proprietary and either not published or expensive to access, which makes it difficult for reviewers and readers to evaluate item quality and the validation techniques that were used to support the use of the scale. This would be less of a concern if contemporary best practices were routinely followed when constructing and validating surveys, but this is not necessarily the case (this will be discussed more in the following sections).

Notedly, the validation articles that used CFA would often test different structural models, relying on model fit to adjudicate which psychometric model should be retained. If any one of the psychometric models fit well according to Hu and Bentler’s cutoffs, we noted that the authors typically concluded that the survey had good psychometric properties and was therefore a reliable and valid measure (see, e.g., Crawford & Henry, 2003). However, if the psychometric structure does not replicate across samples, this may actually be evidence of invalidity of a scale and a contributing factor to psychology’s replication crisis. This is because the proper scoring of a survey is dictated by the structure of the latent variable as determined by the fit of the model. Inconsistency in factor structure results in inconsistency in survey scoring, leading to ambiguity of score meaning and irregular relationships with external criterion (McNeish & Wolf, 2020). For example, Duckworth et al. (2021) note that their Grit scale has been modeled by different authors as a higher-order multi-dimensional model, a unidimensional model, and a bi-factor model, causing them to conclude that factor analysis should not be used to solve theoretical problems. Conversely, without clarity about factor structure and item quality, and without consistency in scoring and score meaning, survey use will exacerbate theoretical problems in psychology. Perhaps qualitative evidence (such as item interpretability and relevance) should be collected to better understand *why* the factor structure does not replicate instead of abandoning psychometrics all together due to unfavorable outcomes. Without rigorous validation efforts, it is impossible to identify which scales should be used and which should be revised or abandoned.

That said, Hu and Bentler's approach to identifying model fit is not necessarily the best one; there are many other ways to identify misfitting models (e.g., investigating local areas of misfit or using different types of misspecifications in the data generating model). For most models, the DFI algorithm can be standardized and extended to reconceptualize misfit as a continuum from trivial to substantial, which is considered to be a better practice (McNeish & Wolf, 2021, 2021). DFI cutoffs based on Hu and Bentler's approach to misspecification were only used in this review because, given the popularity of their paper, the field of psychology seems to have implicitly agreed upon this standard. Thus, the goal of this review is not to conclude that these scales are suddenly invalid because they did not meet the revised Hu and Bentler cutoffs or to encourage psychologists to abandon psychometric models. Instead, our intent is to encourage researchers to pause when considering the validity evidence presented to them. Psychology and other social sciences have long relied on Hu and Bentler's cutoffs as a necessary threshold of model fit despite warnings that they should not be applied broadly across all model subspaces (see, e.g., Hancock & Mueller, 2011; Hayduk, 2014; Heene et al., 2011; Hu & Bentler, 1999; Marsh et al., 2004; McNeish et al., 2018; Saris et al., 2009). In seeing that relying on conventional yet outdated approaches to validation may change the inferences about the strength of validity claims, our hope is that researchers will be willing to revisit approaches to validation.

Validity and Validation

Validity and validation have evolved substantially since the origin of psychological measurement in the early 1900s. Validity was first defined in 1921 by the Standardization Committee of the North American National Association of Directors of Educational Research as the degree to which a test measures what it intends to measure (Newton & Shaw, 2013). However, in the late 1900s, it became apparent that test results could not be interpreted absent of ever-evolving social contexts, leading psychometricians to conclude that it is essentially impossible to prove that tests themselves are definitively valid or that they measure what they intend to measure (Kane, 1992; Messick, 1990; Sireci, 2007). As such, validity was reconceptualized as the quality or trustworthiness of the *use* of inferences drawn from self-report survey results and validation was considered to be an on-going process. Nonetheless, it is difficult to update a century of established tradition, and social scientists may find that psychometric coursework on validity theory (when included in introductory textbooks) is "always outdated and often flawed" (Borsboom, 2006, p. 436). As such, the evidence presented in support of the survey use today is often incomplete and antiquated (Cizek et al., 2008; Hubley et al., 2014; Slaney, 2017).

The first set of guidelines for the development and validation of psychological tests and self-report surveys was published in 1952 by the American Psychological Association. It is now known as the *Standards for Educational and Psychological Testing*; the most recent edition of which was published in 2014. Today, professional standards and guidelines about validity and validation exist in at least seven professional subject areas (Chan, 2014)[17]. While the naming of sources of validity evidence do not overlap perfectly across disciplines, there are common themes and expectations that should be familiar regardless of the domain of

[17] The seven professional organizations and their corresponding published guidelines referenced here are (1) Standards for Educational and Psychological Testing (AERA et al. 2014), (2) Guidance for Industry – Patient-Reported Outcomes Measures: Use in Medical Product Development to Support Labeling Claims (Food and Drug Administration, 2009), (3) Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN; Mokkink et al. 2010), (4) Evaluating the Measurement of Patient-Reported Outcomes (EMPRO; Valderas et al. 2008), (5) Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, 2003), (6) Test Reviewing for the Mental Measurement Yearbook at the Buros Center for Testing (Carlson and Geisinger 2012), and (7) European Federation of Psychologists' Association's (EFPA) review model (Evers et al. 2013).

origin. The most commonly recommended sources of validity evidence spanning the fields of business, education, health, and psychology are the content of the survey items, the relationships between the survey and external criterions, and the extent to which survey items relate to each other (Chan, 2014). Along with these sources of validity evidence, metrics of reliability (or data quality) are recommended as well. The division of validity into this “holy trinity” of evidence harkens back to the conceptualization of best practices in validation for most of the 20th century (Cronbach & Meehl, 1955; Guion, 1980; Messick, 1975; Newton & Shaw, 2014). Many social scientists may recognize this triad as content validity, criterion (convergent/predictive) validity, and construct validity, which dominated mainstream validity theory for nearly 50 years (Slaney, 2017).

Beginning with the 1999 edition of the *Standards*, psychometricians have recommended five sources of validity evidence that could be presented in support of the use of a self-report survey. These are evidence for validity based on the test content, the response process, the internal structure, relationships to other variables, and the consequences of survey use (AERA, APA, NCME, 2014). Notedly, the type of evidence that should be presented depends upon the intended uses of a survey and the validity argument one wishes to make, which should be combined into a holistic argument about the quality of the survey (Kane, 2013). Three of these sources of evidence should sound familiar: content validity maps onto evidence based on the test content, construct validity evidence is similar to evidence based on the internal structure, and criterion validity is evidence based on relationships to other variables. The inclusion of response process evidence (i.e., the cognitive process participants engage in when responding to items) was introduced by Embretson (1983), while evidence based on the consequences of survey use (i.e., the ethical application of survey results to make decisions) was introduced by Messick (1975). The fact that these two types of validity are necessary to include speaks to the inherent difficulty of measurement in the social sciences in comparison to the physical sciences, as it would not be similarly necessary or meaningful to evaluate the consequences of creating, for example, a thermometer.

Evidence supporting relationships to other variables, the internal structure, and reliability are typically presented quantitatively, while content, response process, and consequences likely require substantiation that is more qualitative in nature. Relationships to other variables are typically modeled using correlation coefficients or regression coefficients (e.g., a scale designed to measure subjective wellbeing may be correlated with a scale about life satisfaction to examine if they are measuring similar topics). Internal structure evidence is typically presented using a latent variable model such as CFA or item response theory (IRT), both of which assume that the latent variable is continuous in nature (i.e., that the property of interest is a homogeneously ordered continuity; Michell, 2012). Reliability is typically estimated using internal consistency measures such as coefficient alpha or consistency over multiple test administrations (i.e., correlating the same test at two time points). Test content evidence can be presented by mapping test items to the content domain, which can be judged by experts or the population of interest. Response process evidence is typically collected using cognitive interviews or think alouds. Consequences of testing refers to the soundness of the stated intended uses and interpretations of of an assessment; in this case, researchers may collect qualitative or quantitative evidence to demonstrate that the survey use results in realized benefits and does not lead to unintended negative outcomes for participants (AERA, APA, NCME, 2014; Messick, 1975). The *Standards* states that none of these sources of validity evidence are inherently more important than the other, and the type of evidence presented depends on the validity argument one wishes to make (Kane, 1992, 2013).

Sources of Validity Evidence

Nonetheless, the sources of validity evidence that can be demonstrated quantitatively appear to dominate the literature. Quantitative evidence of validity is valuable, but as demonstrated in this study and in others mentioned below, it cannot be relied upon as the sole source of validity evidence. This is primarily because satisfactory results can still be found absent of measurement or meaning. For example, Maul (2017) administered nonsensical and blank survey items and found favorable psychometric properties after running an exploratory factor analysis model (e.g., high factor loadings, coherent factor structure, and high reliability coefficients). Hayduk (2014) demonstrated that causally misspecified CFA models can have non-significant chi-square values due to a phenomenon known as equivalent model fit. Similarly, Rhemtulla et al. (2020) showed that it is possible to successfully fit a common factor model to data that were generated under an entirely different structure. Arnulf et al. (2014, 2018) revealed that it was possible to accurately predict people’s responses to a survey given only semantic information about the survey items, indicating that covariation between items could occur for reasons other than relationships between attributes. Borsboom et al. (2004) and Zumbo (2009) remind us that relying on correlational evidence of validity (via the nomological net; Cronbach & Meehl, 1955) is often theory-avoidant, in that it offers evidence of validity only by confirming or refuting the existence of relationships to other variables or constructs and thus does not *explain* the relationship between those attributes or reference the ontology of the attributes in question. Indeed, a popular example of confounding variables in introductory psychology courses is the statement that eating ice cream causes people to die from shark attacks, prompting professors to conclude that "correlation is not causation"[18].

There have been several other meta-analyses that summarize the types of validity evidence that are most commonly presented; their findings are similar to the results of this review. Cizek et al., (2008) reviewed 283 mental assessments from the 16th edition of the Mental Measurements Yearbook (MMY) and found that 76.3% reported a reliability coefficient, 67.2% reported criterion-related evidence, 58% reported results from a factor analysis, 48.4% reported general references about the test content, and 10.6% reported face validity evidence, while only 2.5% reported evidence about the consequences of testing and 1.8% reported response process evidence. Hubley et al. (2014) reviewed 50 articles from two psychological assessment journals published between 2011 – 2012, and found that 90.2% reported a reliability coefficient, 76.8% reported criterion-related evidence, and 73.2% reported evidence based on the internal structure (factor analysis or measurement invariance), while only 1.8% reported response process and 0% reported content evidence or consequences evidence. Meanwhile, Barry et al. (2014) reviewed 967 articles in health education and behavior between 2007 and 2010, and found that 49% reported a reliability coefficient, 29% reported content evidence, 26% reported internal structure evidence (mostly factor analysis), 12% reported face validity, and 7% reported criterion-referenced evidence (while response process and consequences of testing were not mentioned). Flake et al. (2017) reviewed 500 scales from 122 articles in social and personality psychology in 2014, and

[18] While ice cream consumption and shark attacks are highly correlated, it is because they both happen more often when the weather is warm.

found that 75-80% reported a reliability coefficient (mostly alpha), while 2.4 - 20.9% reported results from a factor analysis model (likewise, the authors did not investigate other types of evidence of validity). In general, results from meta-analyses show that the frequency in which quantitative evidence of validity (e.g., coefficients of validity and reliability) is presented is increasing over time while qualitative evidence is rarely reported, and there is a disconnect between contemporary validity theory and validation (Borsboom, 2006; Chinni & Hubley, 2014).

Quantitative evidence of validity has been popular since the evolution of psychometric testing in the early 1900's (Slaney, 2017). It is possible that quantitative evidence is reported more often today because it is simple and concise, and much faster to collect and synthesize than qualitative data, which is valuable in an academic culture characterized by "publish or perish" (Castillo-Díaz & Padilla, 2013; Launeanu & Hubley, 2017). It is also possible that emphasis on quantification is circular, i.e., it is what is seen most often and therefore it is expected to be included (e.g., some reviewers might gatekeep publishing by requiring a reliability coefficient; Maul, 2017). The absence of clear standards of how to investigate or report qualitative evidence of validity likely contribute to its scarcity in the literature (Hilton, 2017; Presser et al., 2004). Nonetheless, the current best practices in psychometrics state that all types of validity evidence are equally important and that the type of evidence that should be presented depends on the validity argument one wishes to make (AERA, APA, NCME, 2014; Kane, 2013).

Evidence Based on the Response Process

If a researcher wishes to make an argument that survey items were understood as intended and that participants use cognitive processes related to the psychological attribute in question to respond to items designed to measure that attribute, they may be inclined to present evidence for validity based on the participant's response process. Introduced in the 1999 edition of the *Standards*, the participant's response process can be thought of as the cognitive process that an individual engages in when responding to an item on an assessment (Padilla & Benítez, 2014). Although it is rarely included in validation studies, it has been described as essential and as validity in its own right (Launeanu & Hubley, 2017; Wilson, 2005). Evidence of item interpretability is often collected qualitatively, either through cognitive interviews, focus groups, or through web probing (Behr et al., 2017; Peterson et al., 2017; Presser et al., 2004; Willis, 2005). This source of evidence is often described as "pretesting" in which items are tested for comprehension and relevance before being administered to a larger sample (Hilton, 2017).

Though these methods provide valuable insights about item interpretability, there are no clear standards for reporting results or frameworks for re-testing revisions of items that were not understood as intended (Hilton, 2017; Willis, 2005). To remedy this shortcoming, Wolf et al. (Paper 1; 2022) recently introduced the Response Process Evaluation (RPE) Method. The RPE method uses web probes to conduct condensed cognitive interviews known as meta-surveys, focused on participant item comprehension and justification for response option selection. The meta-surveys are administered in small batches of five participants at a time, allowing researchers to get quick feedback about item interpretability, revise items that are not understood as intended, and immediately test the revised items in a new sample of participants. These insights are summarized in a standardized validation report in which researchers document the final version of the item, its intended interpretation, the population for which it was validated, the percent of participants that understood the item as intended, examples of participant interpretations, and any common misinterpretations to be cautious of.

In using the RPE method to revise and validate items, researchers are essentially adopting a bottom-up approach to item development in which the population of interest can contribute to the development of each survey item. To code participants' paraphrases and feedback, researchers must be clear about the definition of the property they intend to measure, the intended use of the survey, the reason each item is included, and the intended interpretation of each item, all of which is recommended by the *Standards* but rarely considered in contemporary scale construction. This thoughtful approach to item development would be a useful compliment to Wilson's (2005) construct mapping approach to creating surveys. Additionally, such validity evidence could be a valuable source of validity evidence for the individual items (or nodes) used in psychological network models (Epskamp et al., 2018), in ecological momentary assessment (EMA), or with single-item scales. Evidence that participants interpret the survey items as intended does not have to be presented, but researchers should be aware that omitting evidence of item interpretability essentially makes the claim that item interpretability is not important to test.

Reincorporating Theory into Psychometrics

Ideally, prior to constructing a scale, psychologists would begin by clearly defining the attribute to be measured, articulating the intended uses of the scale, hypothesizing the structure of the attribute (e.g., categorical or continuous) and specifying the hypothesized relationship between the attribute and the responses to the items (e.g. parametric or non-parametric). In other words, theory would guide the choice of the psychometric model to be used (Borsboom, 2006; Maul, 2017). They might additionally engage in thoughtful item development practices such as construct mapping, i.e., hypothesizing what individuals with different severities or abilities of the property of interest might look like and creating items that measure different levels of that property (Wilson, 2005). Further, they might invite the population of interest to contribute to the development of the survey by pretesting items for coherence and interpretability before administering them to a larger sample; a long recommended but often neglected practice. Additionally, they might consider the consequences of misusing a scale and clearly articulate how it should not be used.

However, for over 120 years, psychologists have operated under what Michell (2003) refers to as the "quantitative imperative", the idea that to be taken seriously as a science, psychologists must be able to measure psychological properties, and for measurement to exist, psychological properties must be continuous (Slaney, 2017). As such, psychometric models have been applied not because they matched a theory that a researcher wanted to test, but because researchers believe that they must test theories using psychometric models (in particular, reliability coefficients from classical test theory and factor analysis, which appear in most papers on instrument validation). Thus, it is reasonable to suggest that psychologists have been designing psychological theories to fit traditional psychometric models instead of using a theory-driven approach to psychological measurement. Further, they may have been neglecting qualitative evidence not because they want to, but because quantitative evidence seems more rigorous and trustworthy.

Researchers often test different theories of measurement by comparing psychometric models to determine which one fits the data best. However, uncovering the "true" structure of a property therefore may not be as simple as examining statistical model fit. If researchers design their instruments to have good psychometric properties under a standard latent variable model, then the instrument must conform to certain criteria. For example, if a researcher believes that they need to report an indicator of reliability such as coefficient alpha, and that coefficient alpha should exceed .7 to demonstrate good internal consistency, then researchers would be motivated to include many similar items on their scale. Similarly, if a researcher wanted to produce evidence of reliability using Omega or Coefficient H, it would be necessary to 1) use a factor model as evidence of validity and 2) maximize the factor loadings (McNeish, 2018). The quantitative imperative could result in item redundancy and construct under-representation since redundancy will return a high reliability coefficient. Further, if instruments are designed to represent a continuous latent variable then, by design, it may be difficult to find evidence to the contrary. For example, if an individual uses items that were designed to fit a factor model and models them using a mixture model, they might find a set of ordered classes. This does not necessarily mean that the underlying structure of the latent variable would be better conceptualized as continuous, but rather that it perhaps should be *expected* to be continuous because the items were designed to fit a model that assumes a continuous psychological property. In other words, designing an instrument to fit a factor analysis model might suppress the true nature of the construct, inhibiting our ability to use psychometric models to evaluate psychometric theory.

This may be why psychological measures do not necessarily match best practices in psychology and may even inhibit the ability to find statistical significance and/or meaningful effect sizes in intervention studies. As noted by Fried and Neese (2015), there are no clinical tests for psychopathology despite decades of anticipation (Kapur et al., 2012) and treatment efficacy may be obfuscated by improperly summing item responses on instruments under the ubiquitous assumptions of quantity and unidimensionality. For example, one of the most popular evidence-based treatments for mental disorders, cognitive-behavioral therapy (CBT), is designed to disrupt maladaptive thought patterns that *cause* emotional and behavioral responses, which violates the assumption of local independence that is essential for latent variable models. In other words, the therapy that clinical practitioners use to treat mental disorders directly contradicts the instruments that are designed to measure symptom severity. The unconscious reliance on CFA as evidence of validity may be part of the reason that psychology is suffering from a replicability crisis. The ontology of the attribute doesn't match the epistemic framework.

Conclusion

In this paper, we have simply reiterated what others have been saying for decades: relying exclusively on a classic set of quantitative analyses is likely insufficient to justify a claim of validity (see, e.g., Borsboom, 2006; Borsboom et al., 2004; Maul, 2017; Messick, 1995; Michell, 2012; Satchell et al., 2021; Wilson, 2005; Zumbo & Hubley, 2017). This does not mean that psychometric models do not add value or should not be used. Rather, if researchers determine that evaluating the internal structure of a survey would be helpful for their validity argument, thoughtful measurement theory should guide the choice of a psychometric model. Additionally, we hope to bring attention to the fact that validity theory and instrument construction has evolved substantially since the late 1900s, and current guidelines suggest that multiple sources of validity evidence be presented depending upon the intended uses and interpretations of survey results. The quality of inferences drawn from survey results would likely be enhanced by the inclusion of qualitative evidence into instrument development.

References

AERA, APA, NCME. (2014). Validity. In *Standards for Educational and Psychological Testing* (pp. 11–31). American Educational Research Association.

Arnulf, J. K., Larsen, K. R., & Martinsen, Ø. L. (2018). Respondent Robotics: Simulating Responses to Likert-Scale Survey Items. *SAGE Open*, *8*(1), 2158244018764803. https://doi.org/10.1177/2158244018764803

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PLoS ONE*, *9*. https://doi.org/10.1371/journal.pone.0106361

Barrick, M. R., & Mount, M. K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology*, *44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and Reliability Reporting Practices in the Field of Health Education and Behavior: A Review of Seven Journals. *Health Education & Behavior*, *41*(1), 12–18. https://doi.org/10.1177/1090198113483139

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Behr, D., Meitinger, K., Braun, M., & Kaczmirek, L. (2017). Web probing – implementing probing techniques from cognitive, interviewing in web surveys with the goal to assess the validity of survey questions. *Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_en_023

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. https://doi.org/10.1037/0033-2909.88.3.588

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

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). The Guilford Press.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. https://doi.org/10.1016/0165-1781(89)90047-4

Castillo-Díaz, M., & Padilla, J.-L. (2013). How Cognitive Interviewing can Provide Validity Evidence of the Response Processes to Scale Items. *Social Indicators Research*, *114*(3), 963–975. https://doi.org/10.1007/s11205-012-0184-8

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., … Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Chan, E. K. H. (2014). Standards and Guidelines for Validation Practices: Development and Evaluation of Measurement Instruments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9

Chinni, M. L., & Hubley, A. M. (2014). A research synthesis of validation practices used to evaluate the Satisfaction with Life Scale (SWLS). In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral and Health Sciences* (pp. 229–241).

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement*, *68*(3), 397–412. https://doi.org/10.1177/0013164407310130

Coates, R., Ayers, S., & de Visser, R. (2017). Factor structure of the Edinburgh Postnatal Depression Scale in a population-based sample. *Psychological Assessment*, *29*, 1016–1027. https://doi.org/10.1037/pas0000397

Cox, J. L., Holden, J. M., & Sagovsky, R. (1987). Detection of Postnatal Depression: Development of the 10-item Edinburgh Postnatal Depression Scale. *The British Journal of Psychiatry*, *150*(6), 782–786. https://doi.org/10.1192/bjp.150.6.782

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology*, *42*(2), 111–131. https://doi.org/10.1348/014466503321903544

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2021). Revisiting the Factor Structure of Grit: A Commentary on Duckworth and Quinn (2009). *Journal of Personality Assessment*, *103*(5), 573–575. https://doi.org/10.1080/00223891.2021.1942022

Dunbar, M., Ford, G., Hunt, K., & Der, G. (2000). A confirmatory factor analysis of the Hospital Anxiety and Depression scale: Comparing empirically and theoretically derived structures. *British Journal of Clinical Psychology*, *39*(1), 79–94. https://doi.org/10.1348/014466500163121

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

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), 1–11. https://doi.org/10.1186/s12916-015-0325-4

Galinha, I. C., Pereira, C. R., & Esteves, F. G. (2013). Confirmatory factor analysis and temporal invariance of the Positive and Negative Affect Schedule (PANAS). *Psicologia: Reflexão e Crítica*, *26*, 671–679. https://doi.org/10.1590/S0102-79722013000400007

Gelaye, B., Lohsoonthorn, V., Lertmeharit, S., Pensuksan, W. C., Sanchez, S. E., Lemma, S., Berhane, Y., Zhu, X., Vélez, J. C., Barbosa, C., Anderade, A., Tadesse, M. G., & Williams, M. A. (2014). Construct Validity and Factor Structure of the Pittsburgh Sleep Quality Index and Epworth Sleepiness Scale in a Multi-National Study of African, South East Asian and South American College Students. *PLOS ONE*, *9*(12), e116383. https://doi.org/10.1371/journal.pone.0116383

Gonzalez, O., Canning, J. R., Smyth, H., & MacKinnon, D. P. (2020). A Psychometric Evaluation of the Short Grit Scale: A Closer Look at its Factor Structure and Scale Functioning. *European Journal of Psychological Assessment : Official Organ of the European Association of Psychological Assessment*, *36*(4), 646–657. https://doi.org/10.1027/1015-5759/a000535

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Graham, J. R. (1987). The MMPI: A practical guide, 2nd ed (2nd ed., pp. xvi, 311). Oxford University Press.

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, *11*, 385–398. https://doi.org/10.1037/0735-7028.11.3.385

Hancock, G. R., & Mueller, R. O. (2011). The Reliability Paradox in Assessing Structural Relations Within Covariance Structure Models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Hayduk, L. (2014). Seeing Perfectly Fitting Factor Models That Are Causally Misspecified: Understanding That Close-Fitting Models Can Be Worse. *Educational and Psychological Measurement*, *74*(6), 905–926. https://doi.org/10.1177/0013164414527449

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Hilton, C. E. (2017). The importance of pretesting questionnaires: A field research example of cognitive pretesting the Exercise referral Quality of Life Scale (ER-QLS). *International Journal of Social Research Methodology*, *20*(1), 21–34. https://doi.org/10.1080/13645579.2015.1091640

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation*

Modeling: A Multidisciplinary Journal, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of Validation Practices in Two Assessment Journals: Psychological Assessment and the European Journal of Psychological Assessment. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 193–213). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_11

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*(12), 1174–1179. https://doi.org/10.1038/mp.2012.105

Launeanu, M., & Hubley, A. M. (2017). Some Observations on Response Processes Research and Its Future Theoretical and Methodological Directions. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 93–113). Springer.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Maslach, C., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior*, *2*(2), 99–113. https://doi.org/10.1002/job.4030020205

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

Maul, A. (2018). Validity. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

McCrae, R. R., & Costa, P. T. (2008). The five-factor theory of personality. In *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). The Guilford Press.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

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

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*(10), 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Messick, S. (1990). Validity of Test Interpretation and Use. *ETS Research Report Series*, *1990*(1), 1487–1495. https://doi.org/10.1002/j.2333-8504.1990.tb01343.x

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, *50*(9), 741–749.

Michell, J. (2003). The Quantitative Imperative: Positivism, Naive Realism and the Place of Qualitative Methods in Psychology. *Theory & Psychology*, *13*(1), 5–31. https://doi.org/10.1177/0959354303013001758

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Frontiers in Psychology*, *3*, 1–8. https://doi.org/10.3389/fpsyg.2012.00261

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, *42*(5), 875–881. https://doi.org/10.1016/j.paid.2006.09.021

Muck, P. M., Hell, B., & Gosling, S. D. (2007). Construct validation of a short five-factor model instrument: A self-peer study on the German adaptation of the Ten-Item Personality Inventory (TIPI-G). *European Journal of Psychological Assessment*, *23*, 166–175. https://doi.org/10.1027/1015-5759.23.3.166

Mulaik, S. (2007). There is a place for approximate fit in structural equation modelling. *Personality and Individual Differences*, *42*(5), 883–891. https://doi.org/10.1016/j.paid.2006.10.024

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*(3), 301–319. https://doi.org/10.1037/a0032969

Newton, P. E., & Shaw, S. D. (2014). *Validity in Educational and Psychological Assessment | SAGE Publications Inc*. Sage Publications, Inc. https://us.sagepub.com/en-us/nam/validity-in-educational-and-psychological-assessment/book239005

Padilla, J.-L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26.1*, 136–144. https://doi.org/10.7334/psicothema2013.259

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes. *Measurement and Evaluation in Counseling and Development*, *50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for Testing and Evaluating Survey Questions. *Public Opinion Quarterly*, *68*(1), 109–130. https://doi.org/10.1093/poq/nfh008

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Santor, D. A., Gregus, M., & Welch, A. (2006). FOCUS ARTICLE: Eight Decades of Measurement in Depression. *Measurement: Interdisciplinary Research and Perspectives*, *4*(3), 135–155. https://doi.org/10.1207/s15366359mea0403_1

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Satchell, L. P., Fido, D., Harper, C. A., Shaw, H., Davidson, B., Ellis, D. A., Hart, C. M., Jalil, R., Bartoli, A. J., Kaye, L. K., Lancaster, G. L. J., & Pavetich, M. (2021). Development of an Offline-Friend Addiction Questionnaire (O-FAQ): Are most people really social addicts? *Behavior Research Methods*, *53*(3), 1097–1106. https://doi.org/10.3758/s13428-020-01462-9

Sehn, F., Chachamovich, E., Vidor, L. P., Dall-Agnol, L., Custódio de Souza, I. C., Torres, I. L. S., Fregni, F., & Caumo, W. (2012). Cross-Cultural Adaptation and Validation of the Brazilian Portuguese Version of the Pain Catastrophizing Scale. *Pain Medicine*, *13*(11), 1425–1435. https://doi.org/10.1111/j.1526-4637.2012.01492.x

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher*, *36*(8), 477–481. https://doi.org/10.3102/0013189X07311609

Slaney, K. (2017). *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Sullivan, M. J. L., Bishop, S. R., & Pivik, J. (1995). The Pain Catastrophizing Scale: Development and validation. *Psychological Assessment*, *7*, 524–532. https://doi.org/10.1037/1040-3590.7.4.524

Tarka, P. (2018). An overview of structural equation modeling: Its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, *52*(1), 313–354. https://doi.org/10.1007/s11135-017-0469-8

Vanheule, S., Rosseel, Y., & Vlerick, P. (2007). The factorial validity and measurement invariance of the Maslach Burnout Inventory for human services. *Stress and Health*, *23*(2), 87–91. https://doi.org/10.1002/smi.1124

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Whisman, M. A., Perez, J. E., & Ramel, W. (2000). Factor structure of the Beck Depression Inventory—Second Edition (BDI-ii) in a student sample. *Journal of Clinical Psychology*, *56*(4), 545–551. https://doi.org/10.1002/(SICI)1097-4679(200004)56:4<545::AID-JCLP7>3.0.CO;2-U

Willis, G. B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design* (1st edition). SAGE Publications, Inc.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Taylor & Frances.

Wolf, M. G., Ihm, E., Maul, A., & Taves, A. (2022). Survey Item Validation. In S. Engler & M. Stausberg (Eds.), *Handbook of Research Methods in the Study of Religion* (2nd ed., pp. 612–624). Routledge.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, *67*(6), 361–370. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). IAP Information Age Publishing.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9

Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and Investigating Response Processes in Validation Research*. Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5

Conclusion

I had two objectives in this dissertation. The first was to highlight the discrepancy between the theory of validity and the practices that researchers typically engage in to validate survey instruments (Borsboom, 2006; Chinni & Hubley, 2014). There is a wealth of literature that better articulates the shortcomings of some of the most popular approaches to validation (see, e.g., Borsboom et al., 2004; Fried & Nesse, 2015; Maul, 2017; McNeish et al., 2018; Michell, 2012; Slaney, 2017; Zumbo & Chan, 2014); each of which motivated my interest in this topic and caused me to think differently about psychological measurement. Secondly, with the help of my committee members, I suggested some pragmatic, actionable tools to hopefully improve the current state of affairs.

In the first paper, I introduced a new framework for presenting evidence based on the participant's response process. The Response Process Evaluation (RPE) method turns cognitive interviews into meta-surveys through the use of web probes, allowing researchers to collect feedback about item interpretability and follow the participant's response process as they select a response option. This allows researchers to determine if participants are using the expected cognitive processes when responding to items, i.e., if variation in the attribute causes variation in the item response (Borsboom et al., 2004). The RPE method is unique because it introduces an iterative structure for testing multiple versions of items and compiles the results into a standardized item validation report in which researchers can document the intended interpretation of the item, the population for which it was validated, the percent of participants that understood the item as intended, and any common misinterpretations to be wary of. I demonstrated the use of the RPE method on an item taken from the Inventory of Non-Ordinary Experiences (INOE; Taves et al., in preparation).

In the second paper, I wrote a tutorial for a Shiny app and R package that I recently co-created (Wolf & McNeish, 2020, 2022). This software automates the Dynamic Fit Index (DFI) approach to calculating approximate fit index (AFI) cutoff values for confirmatory factor analysis (CFA) models (McNeish & Wolf, 2021, 2022). Briefly, researchers have long relied on Hu and Bentler's (1999) fixed cutoff values to distinguish between correctly specified and incorrectly specified factor models despite the fact that they are not generalizable (Hancock & Mueller, 2011; Heene et al., 2011; Marsh et al., 2004; McNeish et al., 2018; Saris et al., 2009). To accurately identify misfitting models and quantify the degree of misfit it is necessary to calculate model-specific cutoffs, which requires creating a bespoke simulation for each factor model (an arduous task for most applied researchers). To alleviate this burden and make custom cutoffs accessible to all, we created the DFI algorithm and software. User feedback indicated that the software was a little difficult to use and the DFI algorithm was difficult to understand, which motivated the creation of this simple tutorial. We were recently awarded an IES grant to extend the DFI algorithm to other model subspaces (e.g., models with categorical outcomes, higher-order models, and models with non-normal data, among others).

In my third paper, I discussed the problems with over-relying on quantitative evidence of validity to assess the quality of survey instruments in the social sciences. I began by conducting a meta-analysis of some of the most popular scales in psychology and reviewing the types of validity evidence that was presented for each in the original publication. Next, I located articles that used CFA as evidence of validity for these scales and summarized how often the hypothesized factor structure was retained. I subsequently used the Shiny app from the second paper to calculate DFI cutoffs for each model and reported how often the model met the revised cutoffs that were tailored to the user's specific model. I concluded the paper with a discussion about the implications for the practice of validation, encouraging researchers to be skeptical of scales that rely primarily on quantitative evidence and to incorporate qualitative evidence into scale development.

I believe that there are several ways that each of these papers could be improved. In paper one, I demonstrate how the RPE method can be used for a survey item with a binary response option (i.e., yes/no). However, researchers commonly use polytomous response scales (e.g., likert scales) when creating survey instruments. Researchers may need to use additional probes with these types of items. Thus, this paper could be strengthened with the addition of a section about polytomous response options (both categorical and ordered). In paper two (DFI tutorial), I had hoped to include a demonstration of how to write up results after modifying the model (e.g., by adding a residual correlation or testing an alternative factor structure). However, this was not possible because the misspecification for the multi-factor model is not standardized in the same way that the one-factor model is, making the cutoffs for multi-factor models more unpredictable. I hope to soon work on a standardized extension of the multi-factor model so that I can add this to the tutorial. Additionally, I think that it would be helpful for researchers to see an example of a model statement for a multi-factor model.

I have been thinking about the content in the third paper throughout most of my graduate career. I hoped it would be easier to write, but I still found it difficult to articulate some of the points that I wanted to make during the theoretical discussion in the second half of the paper. Thus, there are several ways in which I believe this paper could be strengthened before being published. The first is with the addition of co-authors that have more experience writing about these sorts of issues. On a more basic level, I think it could also be improved by lowering the threshold for the number of citations for the meta-analysis so that more scales can be included (especially scales that were published more recently). It may also benefit from a paragraph about the structure of latent properties (e.g., continuous vs categorical), the assumptions associated with latent variable modeling (e.g., local independence), and how relying on latent variable modeling may shape the way researchers think about the structure of psychological properties. Lastly, a brief discussion about item response theory (IRT) and the consequences of testing could be beneficial.

I learned a lot writing this dissertation, and I hope that each of my papers offers a robust first draft for what I hope will become three published journal articles. Like the current state of psychological measurement, I believe that there is room for improvement. With the help of co-authors and collaboration, I am optimistic that we will be able to offer several serviceable tools to help researchers better construct quality survey instruments.

References

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

Chinni, M. L., & Hubley, A. M. (2014). A research synthesis of validation practices used to evaluate the Satisfaction with Life Scale (SWLS). In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral and Health Sciences* (pp. 229–241).

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

Hancock, G. R., & Mueller, R. O. (2011). The Reliability Paradox in Assessing Structural Relations Within Covariance Structure Models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2022). Dynamic fit index cutoffs for one-factor models. *Behavior Research Methods*.

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Frontiers in Psychology*, *3*, 1–8. https://doi.org/10.3389/fpsyg.2012.00261

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. https://doi.org/10.1080/10705510903203433

Slaney, K. (2017). *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Wolf, M. G., & McNeish, D. (2020). *Dynamic Model Fit* (1.1.0) [R Shiny]. https://www.dynamicfit.app

Wolf, M. G., & McNeish, D. (2022). *dynamic: DFI cutoffs for latent variable models* (1.1.0) [R]. https://cran.r-project.org/web/packages/dynamic/index.html

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54). Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9