Principles and assumptions of psychometric measurement

Gavin T. L. Brown

The University of Auckland, New Zealand http://orcid.org/0000-0002-8352-2351 gt.brown@auckland.ac.nz

Citar como: Brown, G. (2023). Principles and assumptions of psychometric measurement. *Revista Digital de Investigación en Docencia Universitaria, 17(2)*, e2022. https://doi.org/10.19083/ridu.2023.1834

Recibido: 13/12/21. **Revisado:** 24/02/22. **Aceptado:** 21/06/22. **Publicado**: 30/06/2023.

Abstract

Introduction: the validity of claims about learner self-reports depends on the credibility of the measures used. Inventories developed within the psychometric tradition are expected to provide theoretical and empirical evidence for the validity and reliability of the measures to support subsequent interpretations and decisions. These practices depend on a set of assumptions based on latent trait theory that are essential to understanding what psychometric measures can do. This tutorial outlines essential characteristics of design and reporting of psychometric self-report data.

Keywords: self-report; psychometric methods; validity; reliability; evaluation

Principios y supuestos de la medición psicométrica

Resumen

Introducción: la validez de las afirmaciones sobre los autoinformes depende de la credibilidad de las medidas utilizadas. Se espera que los inventarios desarrollados dentro de la tradición psicométrica proporcionen evidencia teórica y empírica de la validez y confiabilidad de las medidas para respaldar las interpretaciones y decisiones posteriores. Estas prácticas dependen de un conjunto de suposiciones basadas en la teoría del rasgo latente que son esenciales para comprender lo que pueden hacer las medidas psicométricas. Este documento describe las características esenciales del diseño y el informe de datos de autoinforme psicométrico.

Palabras clave: autoinforme, métodos psicométricos, validez, fiabilidad, evaluación

Introduction

Decision-making about instruction, intervention, or treatment requires high-quality information about current status (i.e., strengths and weaknesses), trends in change (if any), and impact of precursor and introduced causal factors. Furthermore, understanding of the merit, worth, or value of the obtained values in the information needs to be exercised so as to lead to appropriate decisions, actions, and consequences (Scriven, 1967). Educational and clinical interventions require robust information from those who are supposed to benefit from those interventions (i.e., students, clients). If we seek

***Correspondencia:**
Gavin T. L. Brown
gt.brown@auckland.ac.nz

e-ISSN: 2223-2516 © Universidad Peruana de Ciencias Aplicadas https://doi.org/10.19083/ridu.2023.1834

to change knowledge, skill, attitude, belief, and behaviours, we need accurate measures of where people are and where they have got to post-intervention (Messick, 1989). As such then, all methods of systematically sampling individuals' skill, attitude, ability, or any other characteristic constitute 'tests' (Cizek, 2020).

Psychometrics, then, is the statistical science behind testing and measuring latent human emotions, thoughts, attitudes, beliefs, opinions, ideas, and knowledge that shape manifest behaviours. Because these phenomena exist behind the eyes, between the ears, and in the viscera of human existence, it is essential that the individual provide this information to the tester/assessor. While this information may be biased through optimistic illusions about oneself (Dunning et al., 2004), there really is no option but to obtain and exploit the information given by the individual about the state of their own feelings, cognitions, and intentions. Unlike research with physical or animal phenomena, the human is capable of knowing and communicating about internal states and traits. While intimate partners (e.g., parents, siblings, children, partners) may have privileged insights into the mind and behaviours of an individual, the person with the potentially most complete and current information is the individual herself. Observable behaviours may give useful insights into the mind and will of an individual; nevertheless, quite contrary beliefs, intentions, or motivations can lead to very similar observed actions. Hence, making sense of what is in the human mind depends on obtaining information from the owner of that mind. Other than exceptional animals that have learned to communicate with sign language, only the human individual seems able to have something to say that others cannot know unless the individual reveals that information. Thus, the reliance of psychometric research on self-reported information is both necessary and inevitable. Nonetheless, the reliability and validity of tools used to test or measure these psychological phenomena has to be demonstrated for the results to be deemed credible.

In this paper, I shall provide a high-level overview of or guide to psychometric theory and principles used to evaluate measurements of human knowledge, beliefs and attitudes, or intentions. Readers interested in detailed understanding of psychometric principles and practices are encouraged to read authoritative handbooks on psychometrics (Rao & Sinharay, 2007), educational measurement (Brennan, 2006), or testing (Geisinger, 2013).

Psychometric Theory & Practice

Latent Theory Assumptions

Psychometric theory rests on the assumption that manifest behaviours are explained or caused by variation in latent psychological (i.e., cognitive, emotional, volitional, intentional, etc.) attributes within the individual (Borsboom, 2005). These can be called thoughts, beliefs, attitudes, or any number of invisible, non-material, but causal factors that consistently generate observable behaviours. This assumption arises out of our inability to read each other's psychological phenomena without the ability to communicate with each other verbally (Corballis, 2002) about these things. That we can talk about psychological phenomena (e.g., feelings, thoughts, beliefs, etc.) and believe that we understand what others are talking about suggests strongly that we think these things exist. Hence, latent trait theory about psychological phenomena appears to be a logical conclusion from the nature of our shared existence.

The origins of statistical or mathematical characteristics of latent trait theory date back to Spearman's application of correlation techniques (e.g., rank-order correlations, correction for attenuation, and estimation of test reliability) to patterns in test scores to infer the idea of general mental abilities ('g') that account for shared variation among scores (Clauser, 2022). General mental ability or intelligence is presumed to explain why students who do well in one subject tend to do well on other subjects. Unsurprisingly, research into the nature of intelligence has advanced to multi-dimensional models and described the causal influence of both genetic and environmental factors (Deary, 2001).

The mathematical model of latent trait theory assumes that the underlying ability behind test-taker performance requires just one assumption. Lord (1953) puts that assumption as: "*the trait or ability under discussion can be thought of as an ordered variable represented numerically in a single dimension*" (italics in original, p. 518). By extension, researchers who measure an individual's underlying orientation toward a specific mental phenomenon assume that the latent cause has ordered unidimensional characteristics so that individuals can be located as to their frequency, intensity, importance, likelihood, or valence for that construct (Allport, 1935). The widespread use of Likert's (1932) summated rating scale in social psychology manifests this assumption that there is latent cause that explains the strength and covariance of responses to prompts, questions, or items.

It has been argued that latent factors have not been proven to be quantitative (Michell, 1999). Because evidence for latent constructs can only be obtained from manifest operations (e.g., answers to questions), research into how individuals perceive themselves or phenomena must always rest on an assumption that the underlying trait is something that can be counted or quantified. That our experiences of the world and our abilities to respond to that world differ is not an assumption but a given. Hence, it is relatively easy to accept that all individuals vary within populations around normative values for such diverse psychological phenomena as the nature of one's love for their own mother, their ability to solve problems, their ability to use language(s), and so on. Consequently, as Michell (2008) notes, latent trait theory research still has to be honest to admit that "At present we do not know whether this hypothesis [quantitative models of psychological entities] is true, but we will assume it recognizing that at some point in the future someone needs to investigate it" (p. 12).

Consequently, measurements of a latent construct are a proxy for cognitive processes which are always psychologically multidimensional (Rupp, 2008). Thus, as Rupp (2008) put it:

> it **may still be truly meaningful**, and not just statistically convenient, to summarize data via a unidimensional latent variable model. The resulting latent variable **might** serve as a more effective and efficient pragmatic communication tool that represents a decidedly conscious compromise between cognitive fidelity, empirical feasibility, and utilitarian practicability. (p. 122; emphasis added)

This shows that our measurement, regardless of mathematical approach, of a psychological construct always presumes that we have used a defensibly valid or theoretically robust set of proxies, which, in accordance with scientific fallibility (Popper, 1945), may turn out to be wrong.

An advantage of the quantitative assumption is that the numeric values, especially when obtained from large populations, will behave like a real-valued random variable so that mathematical manipulations (e.g., correlations, mean, standard deviation, etc.) provide insight into variation within populations as to the latent trait. Choices have to be made as to how the numbers created in measurement are statistically analysed, ranging from classical test theory to item response theory (Embretson & Reise, 2000). Nonetheless, an individual's position relative to a social norm is an important factor, within the theory of planned behaviour, in determining intentionality and behaviour (Ajzen, 1991).

Related to the possibility that latent constructs are not quantitative, we must accept that our measurements do not have true zero values or inherent scale increments. The arbitrary nature of the scales we use to measure psychological phenomena does not invalidate the existence of the trait. For example, intelligence tests tend to set population mean to 100 and standard deviation to 15. This is a convenient way to describe and locate individuals relative to others. The mean value of 100 could just as easily have been 500 or 1000, with a concomitant SD of 50 or 100. The meaning of any score relative to those norm values would allow us pragmatically to plan useful educational or clinical interventions. Hence, the mathematics of latent trait theory when combined with a theoretical framework for the nature and function of that latent trait allow us to understand and respond to needs or strengths.

External Validation

Just because latent theory for psychological phenomena is pragmatic, does not mean that a latent construct matters. Observation of systematic relationships between variation in the latent construct and observable behaviours and outcomes in the real world is needed to establish whether a hypothesised latent trait matters. Evidence from observable behaviours or independently generated outcomes scores overcomes the bias of self-report as the sole source of data. Gold-standard validation of self-reported scores requires comparison with other measures that are theoretically similar or different (i.e., convergence and divergence; Campbell & Fiske, 1959). A self-report score that is highly correlated with a previously developed measure of a related construct provides convergent evidence. In contrast, low correlations with a measure of a completely different construct provides divergent evidence for the meaningfulness of the measure. Likewise, divergence across time or informant may also call into question the validity of a self-report. Ideally, the convergent and divergent measures will avoid the methodological weakness of self-report data, which is the point of the multi-trait, multi-method approach. Data from biometric evidence (e.g., *f*MRI), online behaviour, test scores, attendance, and other physical measurements will show if variation in scores is meaningfully related to (i.e., causally or explanatorily) to constructs that should be sensitive to the construct (Borsboom et al., 2004; Zumbo, 2009).

Modern neuropsychology attempts to find specific organs of the brain that map to psychological processes (e.g., firing of mirror neurons when physically grasping is related to understanding others; Kaplan & Iacoboni, 2006). An *f*MRI study revealed that, when participants were given bogus feedback about performance, brain regions associated with negative affect (i.e., posterior cingulate cortex, the medial frontal gyrus, and the inferior parietal lobule) were activated when norm-referenced feedback was given to low-competence participants and also when criterion-referenced feedback was given to high-competence participants (Kim, Lee, Chung, & Bong, 2010). Further, performance-approach goal scores correlated positively with activation in the negative emotion brain areas during norm-referenced feedback.

A review of EEG studies identified strong emphasis in the negative emotion brain areas around feedback-related negativity (Meyer et al., 2021), but suggested that because multiple brain regions are involved in these processes and the limitations of the EEG method, modifications are needed to understand how feedback relates to the brain. The complexity of the brain can also be seen in how visual representations of familiar objects and people are located in multiple brain locations (Quiroga et al., 2005). While laboratory studies using *f*MRI or EEG may reveal how the brain interacts with the mind, the challenge is how those studies relate to the complexities of functioning in the real-world environments. There are differences in results when organisms are studied in laboratory glass tubes (i.e., in vitro) and when they are released into living populations (i.e., in vivo) (Autoimmunity Research Foundation, 2012); let alone how they might behave in a cyber or simulated environment (i.e., in silica). How well human self-reported perceptions about feedback map to brain activity when anticipating or receiving feedback is still not evaluated.

A well-established approach to establishing validity of measures of a psychological construct is analysis of how individual self-reports relate to outcome measures. This is well established in educational testing systems, such as OECD's Program for International Student Assessment (PISA) surveys. Marsh et al. (2006) showed that self-reported psychological constructs (i.e., self-reported interest, self-concept, and self-efficacy) collected in PISA from >100,000 students in 2000 had statistically significant, but modest (.25 < *r* < .35), relationships with performance, with invariance across 25 different jurisdictions. Two things need to be said about this result. First, the relationships are consistent with theoretical expectations of how these psychological constructs function. Second, the effect is modest, in part because in vivo contexts are so complex and because individuals may have variability in the constructs they endorse. Indeed, cluster analysis of motivational variables within assessment of mathematics performance showed that in almost all of the 12 nations analysed across three waves of data, there were individuals who did not have consistent self-reported scores across the three motivational scores and performance depended more or less on the mix of motivations (Michaelides et al., 2019).

Nonetheless, self-reported scores about one's own psychological phenomena is a fraught domain. Not only do humans suffer from memory problems about their experiences (Schacter, 1999), but they also suffer from ignorance about themselves and their competence (Dunning, Heath & Suls, 2004), in part because being honest about inadequacies or failure may threaten their ego (Boekaerts & Corno, 2005); or among adolescents it may be 'fun' to subvert surveys (Fan et al., 2006). To overcome these threats psychometrics proposes a number of methods outlined below.

Theoretical Grounding

The field of psychometrics has argued extensively about how to establish validity evidence for any measure of psychological phenomena (Kane & Bridgeman, 2022). Prior to Messick (1989), validity tended to be thought of in terms of multiple types (i.e., face, content, concurrent, construct, and predictive). However, contemporary understanding is that validity is a unitary concept best captured as 'construct validity' (Cizek, 2020). Evidence for a degree of validity judgment (e.g., 'preponderance of evidence', 'clear and convincing evidence', 'substantial evidence'; Cizek, 2020, p. 26) is achieved by consideration of the various empirical and theoretical arguments for the proposed interpretation of a measure (Cizek, 2020; Kane, 2006; Messick, 1989).

Validation evidence includes the theoretical and explanatory qualities of the measurement tool or instrument; the stimulus items, prompts, or tasks presented to elicit responses need to be theoretically aligned to expert theoretically informed definitions of a domain that include hypotheses of how the construct will influence responding (American Educational Research Association et al., 2014; Schmeiser & Welch, 2006). This means that considerable effort should have gone into specifying the domain and then developing and testing the coherence of the various stimuli or prompts used to elicit responses from individuals to that theoretical specification. Pilot studies (International Test Commission, 2018), expert judgement panels (McCoach et al., 2013), participant think aloud studies (van Someren et al., 1994), and cognitive interviews (Karabenick et al., 2007), and so on, are used to demonstrate that there is evidence that the instrument has prima facie alignment with what it is intended to measure.

Reports of how that test or battery is administered and the kinds of data collected are essential to give confidence that the protocols are replicable and theoretically in accord with the domain. Evaluations that test the theoretical preferred model against alternative explanations also provide evidence that the scales are sound (Cronbach, 1988). Hence, evaluation of the internal structure of an inventory should include multiple competing alternatives.

Internal Structure

The psychometric industry takes a scientific approach in which data collection should generate consistent patterns of responses amongst individuals in accordance with their varying responses to a phenomenon. A key constraint on measuring psychological constructs is that they are in and of themselves not directly observable; they are latent (Borsboom, 2005). As such multiple indicators for multiple causes (MIMIC) are used to reduce error in estimation of the strength and direction of attitude, belief, or value and to better represent the phenomenon of interest (Jöreskog & Goldberger, 1975). With sufficient samples and theoretically designed measurements, mathematical modelling of MIMIC response patterns (e.g., estimate of internal reliability, factor analysis) is used to create evidence for the structure and dependability of the proposed scales or factors (Haertel, 2006).

Approaches that provide evidence about the pattern of responses include scale reliability estimation which can be estimated in multiple ways. Although most researchers are familiar with Cronbach's (1951) alpha, extensive research indicates McDonald's (1999) omega and Hancock and Mueller's (2001) coefficient H are superior methods for establishing reliability of a set of item. Within the Rasch modeling framework, Wright and Stone's (1999) item separation G can be used to claim that items elicit responses coherent with that version of item response theory. Researchers should be aware that very high reliability results (e.g., alpha > .90) can be obtained by writing items that are almost identical in wording (i.e., have high homogeneity) producing a 'bloated specific' (Cattell & Tsujioka, 1964).

While principal component analysis can identify underlying vectors in a data matrix, psychometric theory, with its emphasis on error, relies on the common factor model to identify shared and unique factors underlying the same data (Bryant & Yarnold, 1995). Conventional criteria exist to guide interpretation of the statistical results (Bandalos & Finney, 2010) so that the quality of evidence for the internal structure of a research or measurement tool can be evaluated (i.e., scales with little or poor evidence can be ignored, while those with robust evidence can be used). Once robust measurement of latent constructs is established, scores for each factor can be derived (DiStefano, Zhu, & Mîndrilă, 2009).

With scale scores, individuals and groups can be distributed by their scores, which allows comparison of one construct to another and the comparison of score differences between groups and over time. These analytic techniques are robustly associated with the field of psychometrics as they address issues of demonstrating statistically that the theoretical expectations have been met. These practices help create an argument for the trustworthiness of the information obtained from humans about themselves and for interpretations and uses of that data.

Replicability

In the spirit of scientific research, replication studies can examine the stability of psychometric properties across samples (Makel et al., 2012). Statistical techniques such as multigroup confirmatory factor analysis allow researchers to establish whether statistical models of how participants respond to an instrument vary according to whether the new sample is drawn from the same population or not. Ideally, the psychometric characteristics of the measurement tool should be within chance when applied in a new sample drawn from the same population. Clearly, non-invariance should be expected when samples are from divergent populations. This does not mean the measurement is broken; rather, it suggests that the measurement works differently, or the construct being measured is different across language, age, culture, prosperity, or educational boundaries. Consider the non-invariance found in *Teacher Conceptions of Feedback* inventory between New Zealand and Louisiana which have very different policy frameworks (Brown, Harris, O'Quin & Lane, 2017). Measurements that are deployed with new samples have greater opportunity to generate validation evidence by overcoming chance artefacts associated with the development of the measurement. Consider the similarity of the *Teacher Conceptions of Assessment* inventory within New Zealand and its lack of invariance across jurisdictions and languages (Brown, Gebril, & Michaelides, 2019).

An unfortunate side effect of emphasis on and the complexity of statistical and mathematical modelling of the internal structure of a measurement is that validating evidence from external measures tends to be overlooked. Clearly, it is harder to collect evidence from independent samples, to ask participants to complete parallel or divergent measures at the same time as a new one, and even more difficult to collect independent behavioural evidence so as to make a strong case that a new measure is not only psychometrically robust but also has theoretical and empirical evidence for the relationship between what the *mind* reported and what can be seen from the outside. Nonetheless, without the statistical and mathematical evidence of how a new measure actually works it will not be possible to test theoretical claims about how humans think, feel, and behave. Furthermore, given the complexity of factors impinging upon performance, the actual effect of any specific perception may be quite small. When effects are small and responsive to environmental conditions, they are inherently hard to replicate (Lindsay, 2015).

Challenges

A fundamental problem within psychology is that everything in the life experiences, environments, and physiology of individuals influences everything they think, feel, believe, say, or do. So, it should not surprise us that the impact of any single psychological constructs should be relatively small because it interact with all other things that also matter. Efforts to isolate and understand important psychological factors in human life has unfortunately led to widespread jingle-jangle (i.e., same words with different meanings or same meanings with different words) in the field (Flake & Fried, 2020). Nonetheless, latent trait theory provides us a way into the mind, heart, and mental representations of individuals.

Good psychometric evidence for measurements of any psychological construct should be able to:

1. Demonstrate fidelity to a theoretically robust description of what the construct is, how it functions, and what it should do;
2. Provide evidence that the proposed operationalisation has prima facie credibility against that theory;
3. Demonstrate robust statistical evidence for the coherence of items against the construct design of a measurement model;
4. Provide evidence that the measurements are reproducible from additional samples;
5. Provide evidence that the measurements produce effects on other measures (including self-reports), behaviours, or outcomes that align with theoretical expectations; and
6. Provide evidence that the construct can be manipulated such that measurement scores change and have the theoretically proposed effects.

Readers may be dismayed at the thought that any single report should necessarily achieve all of these things. However, an excellent example of how these concerns can be addressed is visible in Thielsch and Hirschfeld (2019) which in seven studies provided: a theoretical framework, item set development, statistical demonstration of scale or factor properties, demonstrated test-retest reliability, validated the scales against other convergent and divergent measures, experimentally manipulated scores, and created large-sample norms. Indeed,users of any psychometric measure should expect evidence of this kind before settling on the use of a new tool or inventory.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A Handbook of Social Psychology* (pp. 798–844). Clark University Press.

American Educational Research Association, American Psychological Association, National Council for Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Autoimmunity Research Foundation. (2012). *Differences between in vitro, in vivo, and in silico studies*. The Marshall Protocol Knowledge Base. Retrieved 12 November from http://mpkb.org/home/patients/assessing_literature/in_vitro_studies

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 93-114). Routledge.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An international review*, *54*(2), 199-231. https://doi.org/10.1111/j.1464-0597.2005.00205.x

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

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

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Praeger.

Brown, G. T. L., Gebril, A., & Michaelides, M. P. (2019). Teachers' Conceptions of Assessment: A Global Phenomenon or a Global Localism. *Frontiers in Education*, *4*(16). https://doi.org/10.3389/feduc.2019.00016

Brown, G. T. L., Harris, L. R., O'Quin, C., & Lane, K. E. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research & Method in Education*, *40*(1), 66-90. https://doi.org/10.1080/1743727X.2015.1070823

Bryant, F. B., & Yarnold, P. R. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). APA.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin*, *56*(2), 81-105.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, *24*(1), 3-30. https://doi.org/10.1177/001316446402400101

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Clauser, B. E. (2022). A history of classical test theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 157-180). Routledge.

Corballis, M. C. (2002). *From Hand to Mouth: The Origins of Language*. Princeton University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Lawrence Erlbaum.

Deary, I. J. (2001). *Intelligence: A Very Short Introduction*. Oxford University Press.

DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20). https://doi.org/10.7275/da8t-4g52

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69-106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. LEA.

Fan, X., Miller, B. C., Park, K.-E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, *18*, 223-244. https://doi.org/10.1177/152822X06289161

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. Advances in Methods and Practices in Psychological Science, 3(4), 456 –465. https://doi.org/10.1177/2515245920952393

Geisinger, K. F. (Ed.). (2013). *APA Handbook of Testing and Assessment in Psychology*. American Psychological Association.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Praeger.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking Construct Reliability Within Latent Variable Systems. In R. Cudeck, S. d. Toit, & D. Sorbom (Eds.), *Structural Equation Modeling: Present and Future - A Festschrift in Honor of Karl Joreskog* (pp. 195-216). Scientific Software International Inc.

International Test Commission. (2018). ITC Guidelines for Translating and Adapting Tests (Second Edition). *International Journal of Testing*, *18*(2), 101-134. https://doi.org/10.1080/15305058.2017.1398166

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631-639. https://doi.org/10.1080/01621459.1975.10482485

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Praeger.

Kane, M., & Bridgeman, B. (2022). The evolution of the concept of validity. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 181-203). Routledge.

Kaplan, J. T., & Iacoboni, M. (2006). Getting a grip on other minds: mirror neurons, intention understanding, and cognitive empathy. *Social neuroscience*, *1*, 175-183. https://doi.org/10.1080/17470910600985605

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., De Groot, E., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139-151. https://doi.org/10.1080/00461520701416231

Kim, S.-I., Lee, M.-J., Chung, Y., & Bong, M. (2010). Comparison of brain activation during norm-referenced versus criterion-referenced feedback: The role of perceived competence and performance-approach goals. *Contemporary Educational Psychology*, *35*(2), 141–152. https://doi.org/10.1016/j.cedpsych.2010.04.002

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, *22*, 5–55.

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, *26*(12), 1827-1832. https://doi.org/10.1177/0956797615616374

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*(4), 517-549. https://doi.org/10.1177/001316445301300401

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research. *Perspectives on Psychological Science*, *7*(6), 537-542. https://doi.org/10.1177/1745691612460688

Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, *6*(4), 311-360. https://doi.org/10.1207/s15327574ijt0604_1

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument Development in the Affective Domain: School and Corporate Applications*. Springer.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). MacMillan.

Meyer, G. M., Marco-Pallarés, J., Boulinguez, P., & Sescousse, G. (2021). Electrophysiological underpinnings of reward processing: Are we exploiting the full potential of EEG? *NeuroImage*, *242*, 118478. https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118478

Michaelides, M., Brown, G. T. L., Eklöf, H., & Papanastasiou, E. (2019). *Motivational Profiles in TIMSS Mathematics: Exploring Student Clusters across Countries and Time* (Vol. 7). Springer Open & IEA.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press https://doi.org/10.1017/CBO9780511490040

Michell, J. (2008). Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 7-24. https://doi.org/10.1080/15366360802035489

Popper, K. (1945). *The open society and its enemies: The high tide of prophecy: Hegel, Marx, and the aftermath* (Vol. II). George Routledge & Sons.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102-1107. https://doi.org/10.1038/nature03687

Rao, C. R., & Sinharay, S. (Eds.). (2007). *Psychometrics* (1st ed.). Elsevier North-Holland.

Rupp, A. A. (2008). Lost in Translation? Meaning and Decision Making in Actual and Possible Worlds. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1-2), 117-123. https://doi.org/10.1080/15366360802035612

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, *54*(3), 182-203. https://doi.org/10.1037/0003-066X.54.3.182

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Greenwood.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39-83). Rand McNally.

Thielsch, M. T., & Hirschfeld, G. (2019). Facets of Website Content. *Human–Computer Interaction*, *34*(4), 279-327. https://doi.org/10.1080/07370024.2017.1421954

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The Think Aloud Method: A practical guide to modelling cognitive processes*. Academic Press.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. MESA press.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. Lissitz (Ed.), *The concept of validity* (pp. 65-82). Information Age Publishers.

RIDU / Revista Digital de Investigación en Docencia Universitaria / e-ISSN: 2223-2516