anales de psicología / annals of psychology
2024, vol. 40, nº 1 (january), 163-170
https://doi.org/10.6018/analesps.583991

ISSN online: 1695-2294.
https://revistas.um.es/analesps

Validation of Psychometric Instruments with Classical Test Theory in Social and Health Sciences: A practical guide

José-Antonio López-Pina[1,*], and Alejandro Veas[2]

[1]*Department of Basic Psychology and Methodology, University of Murcia (Spain)*
[2]*Department of Developmental and Educational Psychology, University of Murcia (Spain)*

Título: Validación de instrumentos psicométricos en ciencias sociales y de la salud: una guía práctica.
Resumen: Recientemente se ha incrementado significativamente el número de estudios psicométricos junto a avances estadísticos cruciales para evaluar las evidencias de fiabilidad y validez de los tests y escalas de medida. Dada la importancia de proporcionar procedimientos más exactos tanto en la metodología como en la interpretación de las puntuaciones, los editores de la revista Anales de Psicología proponen esta guía para abordar los tópicos más relevantes en el campo de la psicometría aplicada. Con esta finalidad, el presente manuscrito analiza los tópicos principales de la Teoría Clásica de Tests (e.g., análisis factorial exploratorio/confirmatorio, fiabilidad, validez, sesgo, etc.) con vistas a sintetizar y clarificar las aplicaciones prácticas, y mejorar los estándares de publicación de estos trabajos.
Palabras clave: Estudios psicométricos. Fiabilidad. Validez. Análisis factorial.

Abstract: In recent years, there has been a significant rise in the number of psychometric studies, together with crucial statistical advances for validity and reliability measures. Given the importance of providing accurate procedures both in methodology and score interpretation of tests and/or measurement scales, the editors-in-chief of the journal Annals of Psychology have drafted this guide to address the most relevant issues in the field of applied psychometry. To this end, the present manuscript analyses the main topics under the Classical Test Theory framework (e.g., exploratory/confirmatory factor analysis; reliability, bias, etc.) aiming to synthesize and clarify the best practical applications; and improve publication standards.
Keywords: Pyshometric studies. Reliability. Validity. Factor analysis.

Introduction

In Social and Health Sciences it is common to build instruments to objectively evaluate the degree to which an attribute or a construct is presented. Psychometry has not yet been able to produce a standardized procedure -similar to a rule of physical measurement- to measure attributes proposed in the different theories (Michell, 1999). Consequently, professionals rely on a set of procedures under the umbrella of structural validity, whose underlying mathematical model is a linear function, either to examine: a) the relevance of content of items through detailed analysis of their structure and relationship (statistics) with the rest of items (content validity), b) relationships with other latent attributes which measure the same (convergent validity) or different attributes (discriminant validity), c) the relevance of that attribute given its discriminating capacity according to sociodemographic variables (sex , age, educational level, race, among others), e) relevance of showing the adequacy of experimental treatments (responsiveness), and f) to determine reliability of scores obtained with these ad-hoc built instruments. These procedures are part of the Classical Test Theory (CTT, Lord & Novick, 1968).

The weakness of the CTT to sustain invariant structures across populations involves cross-cultural adaptations to calculate psychometric properties according to population characteristics. Thus, changes in language and/or in characteristics of the target population generate new psychometric validation studies since essential characteristics (reliability and validity) are dependent on variability of scores and length of instruments.

An editorial review of psychometric studies reveals a significant percentage meet standard of validation, according to technical knowledge available when they were carried out. However, some manuscripts still use outdated concepts and techniques which bring rejection from scientific journals. Rejection occurs even if the work is well-founded, addresses a new instrument or provides relevant data (reliability, validity, cut-off scores, scales based on random sampling, among others) of the instrument. Therefore, it seems necessary to update some aspects which should be included in a study of psychometric validation.

Current advances in Psychometry allow identification of standards that can serve as a guide for authors using statistical techniques such as Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), or Item Response Theory (IRT). However, this manuscript is oriented exclusively toward studies performed under the CTT framework (Abad et al., 2011; Crocker & Algina, 1986; Lord & Novick, 1968; McDonald, 1999). Studies conducted under IRT are excluded (Abad et al., 2011; de Ayala, 2009; Fisher & Molenaar, 1995; Nering and Ostini, 2010; van der Linden & Hambleton, 1997), such us Rasch modelling (Bond and Fox, 2015; Wright and Stone, 1979; Wright & Masters, 1982), and nonparametric IRT (Sijtsma & Molenaar, 2002) and mixed Rasch models (Boeck & Wilson, 2004).

When EFA and/or CFA is applied to a matrix of variables according to the assumptions established in these techniques (interval scale for scores, multivariate normal distribution, linearity, homocedasticity and independence of errors), results are usually powerful, and have oriented the evolution of some psychological theories. Nonetheless, when applied

*** Correspondence address [Dirección para correspondencia]:**
Jose A. López Pina. E-mail: jlpina@um.es
(Article received: 15/09/2023; revised: 24/11/2023; accepted: 14/12/2023)

to variables whose measurement level does not meet the basic assumptions of these techniques (e.g., using dichotomous or polytomous items), results are ambiguous, non-replicable and can lead to severe errors in the number and interpretation of factor solutions (Bock & Gibbons, 2010; Brown, 2006; McDonald, 1999).

EFA/CFA are deemed the most appropriate techniques to show the existence of an attribute, and authors should consider aspects such as: a) items with respect to their content, b) establishment of convergent or discriminant validity through the multitrait-multimethod matrix , c) study of Differential Item Functioning (DIF), d) a study of interpretability of scores (norms and cut-off scores), e) a study of sensitivity and specificity, and f) analysis of measurement invariance, are as important as the EFA/CFA to show quality of results.

Based on a multitude of studies and scales, we have developed a guide to assist professionals in utilizing the most recent advancements in the validation of psychometric instruments. These guidelines are presented in accordance with the framework of a scientific study. Subsequently, a formal delineation of the principal methodological prerequisites linked with Classical Test Theory (CTT) is provided to aid researchers in constructing a suitable validity study.

The Practical Guide

1. In the Introduction

An introduction where the instrument is presented directly and focuses exclusively on the psychometric studies used is not the most appropriate procedure to convince future readers on its practical usefulness. Thus, in our view, the introduction must succinctly describe a theoretical framework and specify the usefulness of the instrument in clinical and/or community contexts. In addition, it is essential to include previous psychometric studies on which the instrument has been adapted or validated, reporting main results found.

2. In the method

2.1. Participants

a) Some studies generally use incidental samples by snowball or on-line procedures. These procedures do not ensure the representativeness of the sample; therefore, it is recommended to utilize a sampling method in community samples.
b) Many studies validate psychometric instruments with participants from universities. Generally, results from these participants cannot be extrapolated to the general population, so their use is discouraged, unless validation is exclusively for that specific population.
c) The sample size required for psychometric analysis will depend on type of statistical procedures performed in the study. Two-hundred cases or above are normally sufficient for item analyses (Crocker & Algina, 1986, Jackson, 2001). If an EFA is performed, the sample size should be a function of communality between items (minimum 10 cases per item). Moreover, if CFA is performed on exploratory solution, the sample size must be sufficient so that two random samples can be generated; one for EFA and another to confirm the structure with CFA. A CFA should not be performed on the same sample on which a solution with EFA has been obtained, although the opposite is valid (Brown, 2006).
d) Description of groups must be as precise as possible. Information must be provided on socio-demographic variables (sex, age, level of education, social background, race, among others) and clinical variables (if applicable). It is also useful to provide statistical evidence, parametric or nonparametric on the balance of groups in sociodemographic variables.
e) Since instruments comprise items (tasks in instruments of maximum execution or symptoms in tests of typical execution) it is recommended that, as far as possible, a descriptive analysis (mean, standard deviation, bias, and kurtosis) of items be reported, and its corresponding psychometric analysis (homogeneity/discrimination indices, and optionally reliability indices and/or validity indices), clearly specifying the correlational method used. Special care should be taken in item analysis after checking dimensionality. If the proposed instrument is not unidimensional, item analysis can be performed for each dimension.
f) A forgotten topic in most studies refers to ceiling and floor effects of scores. If 15% of participants or more obtain lowest or highest scores, a floor or ceiling effect exists (McHorney and Tarlov, 1995). The presence of these effects can alter content validity and reliability and validity coefficients, limiting the possibility of detecting important changes over time when the instrument is applied.
g) Outliers can severely affect the results of a psychometric analysis. Since outliers can be kept or deleted, it is advisable to perform psychometric analysis with and without them to study their effect on the instrument's structure.
h) Some authors report missing data. It is important to report on percentage of missing data, as well as treatment done with them (e.g., an imputation method) (Enders, 2004; Schafer and Graham, 2002).

2.2. In procedure

If a new instrument is introduced, it should include all steps: underlying theory, specification table, selection of tasks /symptoms, and inter-rater study (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

If the instrument has required an adaptation/translation, authors must provide a detailed description of the back-translation process employed (e.g., Hambleton et al., 2005; Muñiz and Bartram, 2007) to ensure complete equivalence between the original and the translated instrument. If the instrument has copyright, the informed consent of the author/s must be available to implement the back-translation process. Furthermore, it is necessary to have the approval of the Ethics Committee of the University or the Hospital where the study is conducted.

In addition, this section should fully describe how the instrument was applied and whether informed consent of participants (parents or guardians in the case of groups with minors) is available.

2.3. In instruments

In this section, the instrument must be fully described: test length, item categories and their justification (if test is newly created), expected dimensions (if proceed) and format used for its administration (e.g., self-report, clinical interview, observation of evaluator or computerized administration, among others).

Instruments used to establish concurrent, convergent, and discriminant validity should also be described. Reliability and validity coefficients of these instruments should be presented for the group/s and/or dimension/s. No work will be accepted if instruments used for concurrent, convergent and discriminant validity have not been previously validated in other studies.

2.4. Statistical analysis

This section describes all procedures used to validate the instrument (item analysis, dimensionality, reliability, and validity coefficients). To this end, different procedures could be used, such as evaluating internal consistency, test-retest reliability, inter-rater and/or intra-rater coefficients, concurrent, convergent and discriminant validity, comparison of groups, and psychometric norms, if applicable.

2.4.1. Exploratory Factor analysis

If an EFA is performed specify a) parameter estimation method, b) factor selection method(s), c) in multidimensional solutions, rotation method and its justification, d) cut-off score of factor loadings, and e) percentage of variance explained by each factor. A complete guide on how to carry out an EFA can be found in Lloret-Segura et al. (2014, 2017) and Ferrando & Lorenzo-Seva (2014). It completely discourages a) performing EFA/CFA on Pearson correlation matrices; b) using Principal Components Analysis (PCA) as method of estimating parameters, and 3) using the Kaiser rule (eigenvalue > 1) as a method of selecting factors. A recent methodological review (Goretzko et al., 2021) addressed four major issues and their practical implications related to EFA: sample size, estimation method, rotation method, and factor retention criterion determining number of factors.

Since EFA began, the relationship between the manifest variable and the factor has been popularly known as factor 'loading'. However, in practice, a factor analysis produces two types of loadings: 1) a structure coefficient, and 2) a pattern coefficient, to show the relationship of the item with the factor. The structure coefficient represents the zero-order correlation between item and factor, while the pattern coefficient represents the unitary effect of a factor on the item, assuming that the effects of the rest of items are biased. When the solution is unidimensional, or multidimensional but orthogonal, the structure and pattern coefficients are equal. However, if the solution is multidimensional with related factors, both coefficients are different and must be included in the manuscript. The term "factor loadings" could be replaced by structure coefficients (unidimensional and multidimensional orthogonal solution) and pattern coefficients should be added in multidimensional solutions with oblique factors.

2.4.3. Confirmatory factor analysis

In CFA, authors must make a previous hypothesis on the dimensional structure of the test based on a) previous theory, b) other factor solutions found in psychometric studies, or c) a previous EFA with a different group of participants in the same study. All three options are feasible, considering that: a) the use of CFA is clearly justified by the underlying theory, and not simply an alternative to EFA, b) if CFA is performed to contrast with exploratory results, it must be justified why this analysis rather than a new EFA is being performed, and c) assuming that EFA and CFA are carried out in the same study, the group where CFA is performed must be different from the group where EFA is performed. It is not advisable to conduct CFA and EFA on the same group (Brown, 2006). However, if CFA solution is not satisfactory it is acceptable to perform an EFA on the same sample.

In CFA, the preferred method to estimate factor solutions is the maximum likelihood on variance-covariance matrices where information of the vectors of means and standard deviations is incorporated; items have five categories or more, and the normality assumption is met (Finney and DiStefano (2006).

The maximum likelihood method requires the assumption of normality of distribution of items. It can be used optionally if the appropriate software is not available, and it must be justified that bias and kurtosis of items are $< |2.0|$. What happens when this method is used with indicators with ordinal values and with ceiling and floor effect (lack of continuity)? Brown (2006, p. 387) stated that:

> '…the potential consequences of treating categorical variables as continuous variables in CFA are multiple, including that (1) they can produce attenuated estimates of the relationships (correlations) between indicators, especially when there are ceiling and floor effects; (2) it leads to 'pseudofactors' that are artifacts of the difficulty of the items and

their extremes, and (3) it produces statistical evidence and incorrect typical errors. Maximum likelihood can also produce incorrect estimates of parameters...'

Therefore, it is essential to employ any method other than maximum likelihood with categorical data or with severely non-normal data'. Beauducel and Herzberg (2006) have also argued against the use of the maximum likelihood method in factor analysis of polytomous items (Flora & Curran, 2004; Lei, 2009).

Brown (2006) proposed several methods to perform CFA of items: a) weighted least squares (WLS), b) unweighted least squares (ULS) and c) robust weighted least squares (WLSMV), currently considered the most recommended method for this type of analysis (Beauducel & Herzberg, 2006; Bentler & Yuan, 1999; Flora & Curran, 2004; Forero et al., 2009, Lei, 2009).

Alternatively, authors should consider employing a bi-factor confirmatory analysis (Bock & Gibbons, 2010; Reise et al., 2007) or exploratory bi-factor models (Asparouhov & Muthen, 2009) as tools to determine a general factor and as many specific factors as considered relevant. When using bi-factor models, it is crucial to justify the theoretical assumptions underlying the hypothetical model. Moreover, as these models tend to have better general fit, it is recommended that authors report rival bi-factor models in comparison to oblique first-order and higher-order structures (Canivez, 2016). Reporting Omega-hierarchical and omega-subscale are also advisable, as these measures are more appropriate indicators of proportion of reliable variable attributable to the latent construct (Zinbarg et al., 2006).

CFA studies should use different fit statistics. Although criteria have changed over time, the most common are chi-square/gl > 2, Standardized Root Mean Square Residual (SRMR < .08), Comparative Fit Index (CFI ≥ .95), Tucker-Lewis Index (TLI ≥ .95), and Nonnormal Fit Index (NNFI ≥ .95). Furthermore, it is essential to examine the matrix of standardized residual covariances to identify local areas of misfit masked in the global fit indices (Brown, 2006).

CFA provides the possibility of comparing competitive models. However, some authors only provide information on the proposed model. Authors are advised to use this advantage, testing all justifiable models according to the theoretical model, comparing models with the χ^2 statistic, if models are nested. If the comparison is between non-nested models, the Akaike information criterion (AIC) or its re-scaled versions (ECVI, CAK, CAIC) can be used (Brown, 2006).

2.4.3. Empirical validity

An important topic for a psychometric instrument is to show its validity in contexts applied by correlating scores of the instrument with those obtained in one or more external criteria. An instrument may be valid in one context and not in another (McDonald, 1999). Nevertheless, some authors believe that determining structural validity through EFA/CFA and calculating reliability of scores with the alpha coefficient is sufficient to show that this instrument can be used with guarantees in practice.

Determining the empirical validity of scores is as important as determining their reliability, thus it is recommended authors incorporate validity coefficients with appropriate external criteria as examples of the actual behavior where the instrument is validated.

A wrong practice which is widespread nowadays, is to calculate empirical validity using another instrument meant to measure the same attribute. This correlation is an expression of the extent to which two instruments measure the same attribute but with different items, but it is not clear evidence of the empirical validity of the instrument in applied contexts. Generally, authors incorrectly use this correlation as evidence of convergent validity. However, this coefficient is evidence of concurrent validity.

Some authors often neglect to perform significance tests (parametric or non-parametric) to determine if the scale really discriminates in the population based on socio-demographic and/or clinical characteristics. Establishing the validity of an instrument is crucial for knowing its practical usefulness. If an instrument does not discriminate between groups of the population for which it has been built its usefulness is nil.

2.4.4. Convergent and discriminant validity

Convergent and discriminant validity coefficients were developed in the context of a multitrait and multimethod matrix. Convergent validity is evidenced when correlations between measurements of the same attribute with different methods (heteromethod-monotrait coefficients) are greater than those between different traits with the same method (heterotrait-monomethod coefficients) (e.g., Mearns et al., 2009). The discriminant validity of the instrument will be evidenced through obtaining low correlations between different traits measured by different methods (i.e., heterotrait-heteromethod coefficients), and must be lower than the convergent validity and reliability coefficients (Crocker & Algina, 1986).

2.4.5. Item categories

One forgotten topic in CTT is to investigate the appropriate number of categories in items, since this model assumes that categories are equiprobable and can be set arbitrarily.

When the instrument is adapted from another culture/language, the norm is to use the same number of categories, however if it is a new instrument, researchers should study the appropriate number of categories for items. Thus, it is desirable to provide a pilot study where different categories have been assessed.

Ways to validate the number of categories are as follows: a) use the Rasch model or polytomous variants: the partial credit model (Masters, 1982) and rating scale model (An-

drich, 1978); b) alternatively, other IRT models such as the graded response model (Samejima, 1969) or generalized partial credit model (Muraki, 1990), or c) use bi-factor confirmatory factor analysis (Canivez, 2016). The advantage of using these models is that categories are experimentally tested. Many statistical software can be used to estimate category thresholds in IRT or Rasch modelling; for example: Winstep (Linacre, 2023), Conquest (Adams et al., 2020), Rummfold (Andrich & Luo, 1996), jMetrik (Meyer, 2014), JAMOVI (jamovi project, 2023), and *R*, among others.

2.4.6. Reliability

In a psychometric study it is advisable to calculate reliability of scores for each dimension measured. Normally, reliability coefficients are reported after performing EFA/CFA. Regardless of the procedure used to obtain the reliability coefficient, researchers should consider the following aspects:

a) The phrase 'the reliability of the test' is incorrect. A psychometric instrument is unreliable. Reliability refers to scores in a particular group or purpose (Thomson and Vacha-Haase, 2000; Vacha-Haase, 1998; Sánchez-Meca et al., 2021). The reliability coefficient is the proportion of true variance that can be attributed to variance of empirical scores. It depends on length of instrument and heterogeneity of scores (Lord & Novick, 1968), as well as other specific characteristics of the group (e.g., O'Rourke, 2004).
b) An undesirable practice is to provide a coefficient of reliability for total scores when the instrument measures two or more dimensions. In our opinion, if the instrument is unidimensional, a reliability coefficient can be presented for the total score, but if the instrument has two or more factors, the reliability coefficient will be presented for each dimension, but not necessarily for the total score. When the instrument measures different dimensions, the reliability coefficient for the total score is an estimator of the actual reliability coefficient, but it is not known to what extent each dimension contributes to the total score.
c) Alpha coefficient (e.g., McDonald, 1999) has long been considered the standard for evaluation of internal consistency, especially when only one application of the measuring instrument is made. Furthermore, this coefficient has been misused as evidence of unidimensionality (Cortina, 1993; Green et al., 1977; Henson, 2001; Schmitt, 1996; Shevlin et al., 2000; Streiner, 2003). Viable alternatives to the alpha coefficient for essentially unidimensional tests are McDonald's coefficient ω (1999) based on factor analysis, or Revelle's β coefficient (Zinbarg et al., 2005) based on cluster analysis. Other useful coefficients that relax the assumptions of CTT (τ-equivalents tests, essentially τ-equivalents and congeneric tests) can provide more realistic information on internal consistency (e.g., Feldt-Brennan coefficient and Feldt-Gilmer coefficients) (Gilmer, & Feldt, 1983; Feldt, & Brennan, 1989). An extensive study of applications of the alpha and ω coefficients can be read in Viladrich et al. (2017).
d) In CFA, it is possible to obtain a reliability coefficient for each subtest through procedures developed by Raykov (2001, 2004) that help in overcoming the problems of the alpha coefficient. An example of how to obtain this reliability coefficient can found in Brown (2006, p. 338-345).
e) If the instrument is self-reporting, it is recommended to utilize a Pearson or Spearman coefficient to assess the temporal stability of the evaluated construct. In cases where the measuring instrument is employed by two or more raters (inter-rater reliability) or by one rater on two occasions (intra-rater reliability), it is more suitable to evaluate temporal stability using the intraclass correlation coefficient (ICC). This coefficient enables the derivation of a reliability coefficient by accounting for the variance attributed to systematic differences between raters or applications.
f) It is highly recommended to present the confidence interval (CI) of the reliability coefficient in each factor or for the complete instrument (unidimensional). If the software does not provide CI, it can be constructed in two ways: a) transforming the reliability coefficient to Fisher's Z scores, and then applying the procedure described by Charter (2000), or b) using the procedure designed by Hastkian and Wallen (1976) when using alpha coefficient. Raykov (2002) developed a method to estimate the CI of the reliability coefficient within the CFA network.

2.4.7. Error of measurement

In general terms, few studies provide an assessment of measurement error and its impact on the quality of scores. Too often, researchers forget that the reliability coefficient is an expression of to what extent scores can be reproduced in successive application of the instrument. However, evaluating measurement error is as important as achieving reliability of scores. To calculate measurement error, CTT provides the measurement standard error (SEM, Crocker and Algina, 1986, McDonald, 1999). SEM allows calculating the interval in which scores can vary within an expected range for the same true score. Perhaps a good idea is to incorporate SEM to detect how much a score must change for a clinically significant change to be detectable (Streiner et al., 2015).

2.4.8. Interpretability of scores

The interpretability of scores refers to the degree to which we can assign qualitative meaning to scores obtained with scales (Lohr et al., 1996). The descriptive statistics obtained on the scale (or subscales) that is validated provide substantial information for interpretability of scores, but authors should provide, if possible, these statistics obtained from the use of the scale (or subscales) in groups or sub-

groups (groups with different clinical diagnosis, age groups, depending on sex, educational level, among others) that are expected to differ in the attribute measured depending on application of a specific treatment.

2.4.10. Bias

In current psychometric studies, it is common to assess whether measurement models hold across different populations or multiple occasions (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). This is typical with populations that may be distinguished regarding language, gender, age, or other characteristics. Measurement invariance is a necessary step for addressing research questions, mainly related to the impact of cultural or linguistic phenomena on item or factor parameters.

According to scientific literature, authors should consider practical application of measurement invariance, such as the application of progressively restrictive constraints on sets of model parameters with multiple groups (Byrne et al., 1989). These constraints have been traditionally named as configural invariance (same model specification to each group separately), metric invariance (constraints on factor loadings), scalar invariance (constraints on measurement intercepts), and strict invariance (constraints on measurement error variances). In addition, if researchers have hypotheses about population differences involving the constructs themselves, this can be done by tests of between group invariance of variances, covariances, and/or means of the latent variables. This is a relevant previous step if researchers wish to evaluate mean differences on measures or constructs between groups (Thompson, 2016). When using categorical data, methods for assessing factorial invariance have been described and involve evaluation of parameters including factor loadings, thresholds, and residual variances (Millsap, 2011; Millsap & Yun-Tein, 2004; Svetina et al., 2020). Nevertheless, it is essential to mention that the accomplishment of measurement invariance is not a mandatory previous step to analyze mean comparison across groups. Two main reasons are behind this statement. First, full measurement invariance is almost impossible to achieve in psychological measurement (Rutkowski & Svetina, 2014). Second, estimation approaches for non-invariance (e.g., partial invariance, invariance alignment, Bayesian approximate invariance) provide different assumptions about DIF distribution, and hence, ambiguity in group comparison. Those interested in further exploring the theoretical and methodological framework of this topic can read the study by Robitzsch and Lüdte (2023). Finally, evaluation of item bias has not yet been incorporated into validation studies within CTT, however, authors are advised to consider the use of procedures to evaluate DIF with any of the procedures devised so far (e.g., logistic regression, or Mantel-Haenszel method) (Brown, 2006), though IRT-based procedures can also be used (De Ayala, 2009). These studies bring greater certainty regarding the invariance of factorial solutions, or of parameters estimated in IRT.

Declaration of Interest Statement: The authors report there are no competing interests to declare.
Funding: No.

References

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud* [Measurement in social and health sciences]. Síntesis.

Adams, R. J., Wu, M. L., Cloney, D., Berezner, A., & Wilson, M. (2020). *ACER ConQuest: Generalised Item Response Modelling Software* (Version 5.29) [Computer software]. Australian Council for Educational Research. https://www.acer.org/au/conquest

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. https://doi.org/10.1007/BF02293814

Andrich, D., & Luo, G. (1996). RUMMFOLDss: *A Windows program for analyzing single stimulus responses of persons to items according to the hyperbolic cosine unfolding model.* [Computer program]. Perth, Australia: Murdoch University.

American Educational Research Association. American Psychological Association. National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438. https://doi.org/10.1080/10705510903008204

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186-203. https://doi.org/10.1207/s15328007sem1302_2

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*, 181-187. https://doi.org/10.1207/S15327906Mb340203

Bock, R. D., & Gibbons, R. (2010). Factor analysis of categorical item responses. In M. L. Nering and R. Ostini (Eds.). *Handbook of polytomous item response theory models*. Routledge.

Bond, T. G., & Fox, C. (2015). *Applying the Rasch model; fundamental measurement in the Human Sciences*. Routledge.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

Byrne, B. M., Shavelson, R. J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Canivez, G. L. (2016). Bifactor modeling. In K. Schweizer & C. DiStefano (Eds), *Principles and methods of test construction* (pp. 247-271). Hogrefe.

Charter, R. A. (2000). Confidence interval formulas for split-half reliability coefficients. *Psychological Reports*, *86*, 1168-1170. https://doi.org/10.1177/003329410008600317.2

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. The Guilford Press.

de Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* Springer-Verlag.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement*, *64*, 419-436. https://doi.org/10.1177/0013164403261050

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (pp. 105–146). Macmillan Publishing Co, Inc; American Council on Education.

Ferrando, P. J., & Lorezo-Seva, U. (2014). Exploratory item factor analysis: additional considerations. *Annals of Psychology*, *30*(3), 1170-1175. https://doi.org/10.6018/analesps.30.3.199991

Finney, S. J. & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G. R. Hancock, & R. O. Mueller (Eds.), *A second course in Structural equation modeling* (pp. 269-314). Information Age.

Fisher, G. H., & Molenaar, I. W. (Eds.) (1995). *Rasch models: Foundations, recent developments, and applications*. Springer-Verlag.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466-491. https://doi.org/10.1037/1082-989X.9.4.466

Forero, C., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*, 625-641. https://doi.org/10.1080/10705510903203573

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with part of unknown lengths. *Psychometrika*, *48*, 99-111. https://doi.org/10.1007/BF02314679

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments, and recommendations for good practice. *Current Psychology, 40*, 3510-3521. https://doi.org/10.1007/s12144-019-00300-2

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827-838. https://doi.org/10.1177/001316447703700403

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum Associates.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189. https://doi.org/10.1080/07481756.2002.12069034

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling*, *8*, 205-223. https://doi.org/10.1207/S15328007SEM0802_3

Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity*, *43*, 495-507. https://doi.org/10.1007/s11135-007-9133-z

Linacre, J.M. (2023). Winsteps® (Version 5.6.0) [Computer Software]. Portland, Oregon: Winsteps.com. Available from https://www.winsteps.com/

Lloret, S., Ferreres, A., Hernández, A., & Tomás, I. (2014). Exploratory item factor analysis: A practical guide revised and updated. *Annals of Psychology*, *30*(3), 1151-1169. https://doi.org/10.6018/analesps.30.3.199361

Lloret, S., Ferreres, A., Hernández, A., & Tomás, I. (2017). The exploratory factor analysis of items: guided analysis based on empirical data and software. *Annals of Psychology*, *33*(2), 417-432. https://doi.org/10.6018/analesps.33.2.270211

Lohr, K. N., Aaronson, N. K., Alonso, J., Burnam, M. A., Patrick, D. L., Perrin, E. B., & Roberts, J. S. (1996). Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clinical Therapeutics*, *18*, 979-992. https://doi.org/10.1016/s0149-2918(96)80054-3

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Masters, G. (1982). A Rasch model for credit partial scoring. *Psychometrika*, *47*, 149-174. https://doi.org/10.1007/BF02296272

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LEA.

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research*, *4*, 293-307. https://doi.org/10.1007/BF01593882

Mearns, J., Patchett, E., & Catanzaro, S. (2009). Multitrait-multimethod matrix validation of the Negative Mood Regulation Scale. *Journal of Research in Personality*, *43*(5), 910-913. https://doi.org/10.1016/j.jrp.2009.05.003

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routdlege.

Michell, J. (1999). *Measurement in Psychology: A critical history of a methodological concept*. Cambridge University Press.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515. https://doi.org/10.1207/S15327906MBR3903_4

Muñiz, J., & Bartram, D. (2007). Improving international tests and testing. *European Psychologist*, *12*, 206-219. https://doi.org/10.1027/1016-9040.12.3.206

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59-71. https://doi.org/10.1177/014662169001400106

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge.

O'Rourke, N. (2004). Reliability generalization of responses by care providers to the Center for Epidemiologic Studies-Depression Scale. *Educational and Psychological Measurement*, *64*, 973-990. https://doi.org/10.1177/0013164404268668

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear restrictions. *British Journal of Mathematical and Statistical Psychology*, *54*, 315-323. https://doi.org/10.1348/000711001159582

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, *37*, 89-103. https://doi.org/10.1207/S15327906MBR3701_04

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, *35*, 299-331. https://doi.org/10.1016/S0005-7894(04)80041-8

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcome measures. *Quality of Life Research*, *16*, 19-31. https://doi.org/10.1007/s11136-007-9183-7

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A multidisciplinary Journal*. https://doi.org/10.1080/10705511.2023.2191292

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31-57. https://doi.org/10.1177/0013164413498257

Samejima, E. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, *12*(4), 516-536. https://doi.org/10.1002/jrsm.1487

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353. https://doi.org/10.1037/1040-3590.8.4.350

Schmitt, N., & Kuljanin, G. (2008). Mesurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences*, *28*, 229-237. https://doi.org/10.1016/S0191-8869(99)00093-8

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*, 99-103. https://doi.org/10.1207/S15327752JPA8001_18

Streiner, D., Norman, G., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford.

Svetina, D., Rutkowski, I., & Rutkowski, D. (2020). Multiple-group invariance with categorical outomes using updated guidelines: an illustration using Mplus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal, 27,* 111-130. https://doi.org/10.1080/10705511.2019.1602776

The jamovi project (2023). *jamovi* (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174-195. https://doi.org/10.1177/00131640021970448

Thompson, M. S. (2016). Assessing measurement invariance of scales using Multiple-Group Structural Equation Modeling. In K. Schewizer & C. DiStefano (Eds.), *Principles and methods of test construction* (pp. 218-244). Hogrefe.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6-20. https://doi.org/10.1177/0013164498058001002

van der Linden, W., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. Springer.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69. https://doi.org/10.1177/109442810031002

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Annals of Psychology*, *33*(3), 755-782. http://dx.doi.org/10.6018/analesps.33.3.268401

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Mesa Press.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133. https://doi.org/10.1007/s11336-003-0974-7

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement, 30*, 121-144. https://doi.org/10.1177/0146621605278814