Assessing construct structural validity of epidemiological measurement tools: a seven-step roadmap

Acessando a validade de construto estrutural de ferramentas de medidas epidemiológicas: um roteiro em sete passos

Acceso a la validez de un constructo estructural de herramientas de medidas epidemiológicas: un guión en siete pasos

Michael E. Reichenheim [1]
Yara Hahr M. Hökerberg [2]
Claudia Leite Moraes [1,3]

[1] Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.
[2] Instituto de Pesquisa Clínica Evandro Chagas, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.
[3] Programa Saúde da Família, Universidade Estácio de Sá, Rio de Janeiro, Brasil.

Correspondence
M. E. Reichenheim
Departamento de Epidemiologia, Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro.
Rua São Francisco Xavier 524, 7º andar, Bloco D, Rio de Janeiro, RJ 20559-900, Brasil.
michael@ims.uerj.br

Abstract

*Guidelines have been proposed for assessing the quality of clinical trials, observational studies and validation studies of diagnostic tests. More recently, the COSMIN (*COnsensus-based Standards for the selection of health Measurement INstruments*) initiative extended those in regards to epidemiological measurement tools in general. Among various facets proposed for assessment is the validity of an instrument's dimensional structure (or structural validity). The purpose of this article is to extend these guidelines. A seven-step roadmap is proposed to examine (1) the hypothesized dimensional structure; (2) strength of component indicators regarding loading patterns and measurement errors; (3) measurement error correlations; (4) factor-based convergent and discriminant validity of scales; (5) item discrimination and intensity vis-à-vis the latent trait spectrum; and (6) the properties of raw scores; and (7) factorial invariance. The paper also holds that the suggested steps still require debate and are open to refinements.*

Epidemiologic Models; Validity of Tests; Methodology

Resumo

*Orientações têm sido propostas para avaliar a qualidade dos ensaios clínicos, estudos observacionais e estudos de validação de testes de diagnósticos. Mais recentemente, a iniciativa COSMIN (*COnsensus-based Standards for the selection of health Measurement INstruments*) estendeu essas orientações para instrumentos de aferição epidemiológicos em geral. Dentre as várias facetas propostas para a avaliação concerne a validade da estrutura dimensional de um instrumento (ou validade estrutural). O objetivo deste artigo é estender essas diretrizes. Um roteiro de sete passos é proposto, examinando: (1) a estrutura dimensional postulada; (2) a força de indicadores componentes relativa ao padrão de cargas e erros de medição; (3) a correlação de resíduos; (4) a validade convergente e discriminante fatorial; (5) a capacidade de discriminação e intensidade dos itens em relação ao espectro do traço latente; (6) as propriedades dos escores brutos; e (7) a invariância fatorial. O artigo também sustenta que os passos sugeridos ainda requerem mais debates e estão abertos a aperfeiçoamentos.*

Modelos Epidemiológicos; Validade dos Testes; Metodologia

Introduction

Several guidelines have been proposed for assessing the quality of validation studies of diagnostic tests since the 1990s [1,2]. While publications aiming to discuss the development process of measurement tools mostly referred to the need for scrutinizing reliability and validity, few aimed to standardize the related nomenclature or specify the required methods to assess these properties [3,4,5].

In this context, the COSMIN – *COnsensus-based Standards for the selection of health Measurement INstruments* – initiative has been evolving in the Netherlands since 2006. Its goal has been to establish standards for the methodological quality of studies evaluating the measurement properties of instruments assessing health status. Basically, the COSMIN proposed three evaluation domains: reliability, validity and responsiveness [6]. The validity domain should cover face and content validity; criterion validity, be it concurrent or predictive, and construct validity. The latter should encompass studies using classical hypothesis testing, and studies on the dimensional structure of an instrument (*a.k.a.* structural validity). The need to assess studies reporting cross-cultural adaptation processes to other sociolinguistic settings has been also emphasized.

In common with the other domains considered in the COSMIN initiative, evaluating the quality of validation studies should be grounded on four cornerstones, *viz.*, type of study design, sample size, extent and management of missing data, and the appropriateness of the employed statistical methods. Specific to the evaluation of structural validity, best quality studies would be those using exploratory and/or confirmatory factor analyses based on classical test theory or item response theory [6].

Although these criteria are unquestionably important, to assess the state of the art on the development process of a given measurement tool and ultimately endorse its suitability for use in epidemiological research, there is also a need to understand the empirical representation of the underlying construct in terms of the properties of the component items and related scales [7,8]. Extending the guideline introduced by COSMIN, the present article is an attempt to organize the steps to follow in the process of assessing the dimensional structure of epidemiological instruments. Beyond statistical technicalities, the article aims to discuss particular evaluation criteria, focusing on interpretability of findings from an applied research perspective.

Essentially, the evaluation of the structural validity of a measurement tool consists of corroborating the hypothesized relationship between latent factors and purported "empirical manifests", i.e., the indicators and related scales. Although this process may potentially involve many viewpoints and approaches [9,10,11], our proposal consists of a seven-step roadmap detailed below.

The seven steps

Step 1: corroborating the dimensional structure

Traditionally, the analytical foundation for evaluating an instrument's dimensional structure has been through factor analyses. The related scientific literature customarily distinguishes an exploratory factor analysis (EFA) from a confirmatory factor analysis (CFA) [10,12]. Yet, the tenuousness of the distinction between "exploration" and "confirmation" is noteworthy if one recognizes that the possibility of a true "confirmation" through a CFA is slight and only materializes if the model under scrutiny effectively happens to be completely substantiated. Otherwise, once few anomalies are uncovered, the researcher immediately falls into an "exploratory mode" regardless if thereafter the method employed to re-specify the model remains of a confirmatory type. Both strategies are thus complementary. Note that there is also a connection from a purely statistical stance since a confirmatory type of factor analysis may be regarded as a particular case, nested within the general class of exploratory models [11].

From an applied perspective then, where should one start the corroborating process? Some authors contend that an EFA should be employed as an initial strategy when a scale is being developed from scratch, supposedly when little or no prior empirical evidence exists, and/or when the theoretical basis is insufficient and frail [13,14]. Once discovering a tenable dimensional pattern, this model would then be submitted to a confirmatory-type modelling process, preferably on a new data set to avoid contamination [14]. An appropriate model fit (cf. Webappendix, note 1; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf) and theoretical coherence would tell the researcher where to stop [9,10].

However, it is worth inquiring whether a "frail theoretical basis" is actually sustainable in any process aspiring to develop a new instrument or in which a consolidated measurement tool is being cross-culturally adapted. If not, it may make little sense to "blindly" implement an EFA in or-

der to "discover" the number of component dimensions and related indicators/items. Maybe, it is reasonable to start with a strong CFA that clearly depicts the theoretically-based conjectured dimensionality, with its manifest indicators, in principle, intelligibly connected to the respective putative factors.

If the course taken is to start with a CFA, the researcher then faces three possibilities. One concerns the unlikely situation mentioned above wherein the specified model flawlessly fits the data and is indisputably acceptable. The second possibility is when only a few abnormalities are identified as, for instance, one or two apparent cross-loadings and/or residual correlations as suggested by Modification Indices (MI) and respective Expected Parameter Changes (EPC) (cf. Webappendix, note 2; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf). Usually, these anomalies go together with a moderate degree of model misfit or, at times, even with adjustments within acceptable levels (cf. Webappendix, note 1; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf). Upholding Jöreskog's classification [15], one would then embark on an *alternative* or *competing model re-specification* process, remaining within a CFA-type framework to estimate freely the proposed features until an acceptable model is reached. The third scenario is when a wide array of possible additional features is suggested by the MIs and/or EPCs. These may not only be indicating that there are several cross-loadings or residual correlations (item redundancies) to be dealt with, but maybe that the entire conjectured dimensional structure is ill suited and untenable. Often, the degree of model misfit tends to be conspicuous if there are a number of anomalies suggested in tandem.

Although engaging in further CFA-type analyses is always possible in this situation, it is perhaps best to turn to a fully "exploratory framework", what Jöreskog [15] called a *model generating* process. As mentioned before, it would be ideal if the newly identified models were subsequently tested via CFA on original data sets pertaining the same population domains. Alternatively, the same data could be randomly split so that the "best" fitting and theoretically coherent model would be identified on part of the sample, and this new model then tested in one or more subsamples also drawn from the total sample. This *half-split* procedure may not be optimal, but could be useful as a first step before a proper corroboration on a new data set is pursued.

Note that this exploratory framework does not imply falling back on the traditional common (principal axis) factor models [9,10]. More complex exploratory models have been recently proposed by Marsh et al. [16], which consist of fitting exploratory models within a CFA framework. Called ESEM (Exploratory Structural Equation Model), this Exploratory/Confirmatory Factor Analysis (E/CFA) holds an advantage over the traditional EFA model in so far as it allows relaxing and effectively implementing some of the restrictions the latter imposes. Freely estimating certain parameters enables testing interesting properties that are otherwise only accomplished with a CFA, yet the main gist of an EFA is kept. Notably, in addition to all loadings being freely estimated and the possibility of rotation as in an EFA, item residuals/error correlations (addressed later) may also be freely evaluated here, which clearly offers more flexibility [17]. Recent developments have reached out to Bayesian frameworks in which EFAs and E/CFAs (ESEM) may also be fit, thus further enhancing tractability [18,19].

Step 2: evaluating item loading pattern and measurement errors

Irrespective of the type of model employed, the scrutiny of factor loadings is implicit in the procedures outlined above since the quest for and uncovering of a sustainable dimensional pattern presupposes an adequate configuration of the item-factor relationship. Several points need assessing, for one, the magnitude of all loadings. Understanding that a completely standardized loading λ_i is the correlation between an item and the related factor, it may be interpretable as representing the strength with which the empirically manifested item expresses signals from the underlying latent trait. Thus the anticipation in a "well behaved" instrument is that all items related conditionally to a given factor show loadings of 0.7 or above. This implies a factor explaining at least 50% of the indicators' variances (λ_i^2). Also known as item reliabilities, these quantities are expressions of the amount the items share with the latent trait, i.e., the *communalities*. The complements of these quantities are the item *uniquenesses* (δ_i), which are properties that should always be reported since they express the amount of information (variance) that remain outside the specified factorial system. Although there is no consensus on the cut-off above which a *uniqueness* is considered high, values above 0.6 should be viewed with some caution, while items with residuals of 0.7 or above could be candidates for suppression and substitution during the measurement tool's development process [9,10,20].

In practice, though, identifying instruments with *all* items showing such "strong" item load-

ings/reliabilities is not that common. Lower loadings may be acceptable, especially if interspersed with others of higher magnitude. For instance, on scrutinizing the pattern of loadings in an exploratory-type model, some authors regard values above 0.3 as tolerable [9,10]. Another suggestion would be to hold values ranging from 0.35 to < 0.5 as "fair" and between 0.5 and < 0.7 as "moderate". These are clearly rough guidelines. Relative loading sizes establishing an overall pattern, and the substantive (theoretical) context will also play a role in aiding the researcher to qualify the pertinence of the item set under inspection.

An *a priori* theoretically-based representation of the instrument's dimensional structure may instruct the researcher as to how items should relate to factors. On a practical level, this most often entails connecting mutually exclusive indicator sets to specific factors. This also needs corroboration since departure from *congenericity* (cf. Webappendix, note 3; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf) – when indicators load on more than one factor – may be regarded as an unwelcome feature [9,21]. Of equal importance, item cross-loadings tend to lower values overall, which in turn implies less than desirable factor-specific item reliability.

Thoroughly examining the pattern of cross-loadings is thus also important. However, uncovering and deciding for the tenability of cross-loadings may not be clear-cut and easy. There are several scenarios to consider, whether to support a cross-loading, or to dismiss it. On fitting a CFA, for instance, highly correlated factors should immediately catch the researcher's eye to the possibility of cross-loadings since the actual modelling solution would be striving for the best adjustment in the light of this unspecified feature. The diagnostic MIs would probably indicate that there is something worth estimating freely, but a high factor correlation in the light of a "proper model fit" may also be flagging undetected cross-loadings. Yet a blatant cross-loading uncovered in an ensuing exploratory-type analysis could be concealing something else, such as an unspecified residual correlation. This would suggest item redundancy needing evaluation from a theoretical interpretative stance (to be covered in the next section).

Regardless of the type of anomaly – whether a small item loading, a cross-loading, or a residual correlation – one possibility would be to eliminate the anomalous items. Ultimately, this would not have many consequences if other items considered appropriate could still map the basic dimension (latent trait). If left unchecked, however, this decision could lead to obliterating content from the latent variable if one or more of its empirical representatives were removed and left out of the scale. This recommendation is all too important in measurement tests implemented as part of cross-cultural adaptation processes, in which originally proposed items are simply discarded because "they are not working" in the new setting. Here, not only content gaps may ensue, but this could also affect external comparability.

Step 3: Examining content redundancy via measurement error correlation

The absence of conditionally correlated items means that items do not share anything beyond the purported latent factor. In evaluating the dimensional structure of an instrument, it is thus important to ascertain this favorable feature or, in contrast, gauge the presence of measurement correlations, which should be unwelcome in principle. Podsakoff et al. [22] provide a thorough overview of possible underlying causes of measurement correlation, which they refer to as common method biases (cf. Webappendix, note 4; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf). Among several conditions, one is of particular interest here, namely, the presence of measurement errors due to at least part of the covariance between items stemming from a common and overlapping stimulus, be it factual or perceived [9].

Conditional independence is a desirable property that should not be assumed *a priori* as often happens, and thus forgo formal assessment. A correlation between residuals may express itself in poor model fit. Inspection is mostly carried out in the context of stringent CFAs through MIs and respective EPCs, but can also be achieved in the context of E/CFAs [16]. However, the actual magnitude of expected parameter changes may not necessarily materialize when freely estimating a residual/error correlation, and thus dismissing the initial suspicion. Sometimes, though, estimated correlations just attenuate and lie within a range that is difficult to interpret, for instance, between 0.3 and 0.5. The decision of what to make of this is not always trivial and as such it may be wise to turn to theory. This brings us to the substantive interpretation of residual (error) correlations.

First, though, there is a need to determine whether the items involved in a given pair load on the same or different factor. If occurring in a congeneric pair (i.e., same factor), the items may truly be semantically redundant. Similar wording is often used unintentionally, even if the aim is to express different underlying ideas or actions.

Other times, repeating content is overtly intended and used to crosscheck a particular point, and items with slightly different phrasing but similar substance are introduced. Regardless, a concern would be if the raw item scores of two highly conditionally correlated items are both used in a total sum of scores. Given that their substantive content overlap, part or most of the content along the latent trait spectrum they intended to cover would end up "doubly-weighted".

Residual correlations require management. Clearly, a solution would be to allow for correlations in any subsequent analysis, but this also entails using intricate models to handle such complexities (e.g., structural equation models). Another solution would be to deal with the items involved. What to do actually depends on the amount of overlap. In the case of very highly correlated items (tending to 1.0), removing one item of the pair – possibly the one with the lowest loading – would be sustainable on the ground that little information would be lost given the almost complete intersection. This situation is uncommon though; most often, residual correlations deserving credit range from 0.3 to 0.6 and the decision to withdraw one of the indicators may be too radical. Information could be lost, especially in scales with few items in which compensation from the other items retained would be less likely. In this case, a viable option would be to join the semantic contents of the two correlated items into a single question designed to capture the information both original items effectively intended to map. Yet this solution, although practical, could run into problems. Some information could possibly be missed by the respondent, depending on the emphasis given to any of the semantic components of the question. In so being, the best solution would be to go all the way back to the "drawing board" and effectively find a single-barreled item as a substitute, and subsequently test its properties in a new study.

Finding a residual correlation between items loading on different factors may also come about. One explanation is that there is a semantic redundancy as perceived by respondents, perhaps due to a dimensional structure misspecification in designing the instrument. In principle, manifests of different putative traits should also hold different contents and semantics. Faulty design notwithstanding, the best solution would be to replace at least one item of the redundant pair. A suggested correlation between errors of items belonging to different factors may be indicative of a dimensional misspecification, especially in the form of an extant cross-loading demanding further exploration. Other possible explanations include pseudo-redundancies caused by other common method variance [22,23]. Once again, resorting to theory may help in resolving this duality.

Step 4: corroborating factor-based convergent and discriminant validity of component scales

Convergent and discriminant validity are properties that have been rather under-appreciated. Convergent validity relates to how much the component items – the empirical manifest – effectively combine in order to map a particular latent trait. In a sense, it captures the joint "communality" of indicators comprising a given factor: as Brown [9] (p. 3) states, "*discriminant (factorial) validity is indicated by results showing that indicators of theoretically distinct constructs are not highly intercorrelated*". In tandem, an instrument is said to hold convergent and discriminant validity if each set of postulated indicators is capable of mapping most of the information on to the related factor in the expected manner, while this amount of information is also greater than that shared across factors (cf. Webappendix, note 5; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf).

The assessment of factor-based convergent validity (FCV) centers on the inspection of the Average Variance Extracted (AVE), which formally gauges the amount of variance captured by a common factor in relation to the variance due to measurement errors of component items (cf. Webappendix, note 6; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf) [10,24]. Values may range from 0 to 1. A factor shows convergent validity if AVE ≥ 0.50, which is indicative that at least 50% of the variance in a measure is due to the hypothesized underlying trait [24]. Seen from the opposite perspective, FCV is questionable if AVE < 0.50 since the variance due to measurement error is then greater than the variance due to the construct [10]. Because it is a summary of what the items supposedly share, lack of factor-based convergent validity is mostly accountable to the influences of one or few component items. Items with weak loadings may contribute little, and a re-analysis without those showing levels of AVE above 0.5 would endorse their removal. Of course, any bolder action would also require a joint appreciation of other features related to the indicator(s) under inspection.

Factor-based discriminant validity (FDV) is also a function of the AVE. A multidimensional model holds FDV when the average variance of each factor is greater than the square of the correlations between this factor and any other factor of the system. For any given factor, the square root of AVE ($\sqrt{\overline{\rho}_{ve(\bullet)}}$) should be higher than the

correlations between this factor and all others in the measurement model.

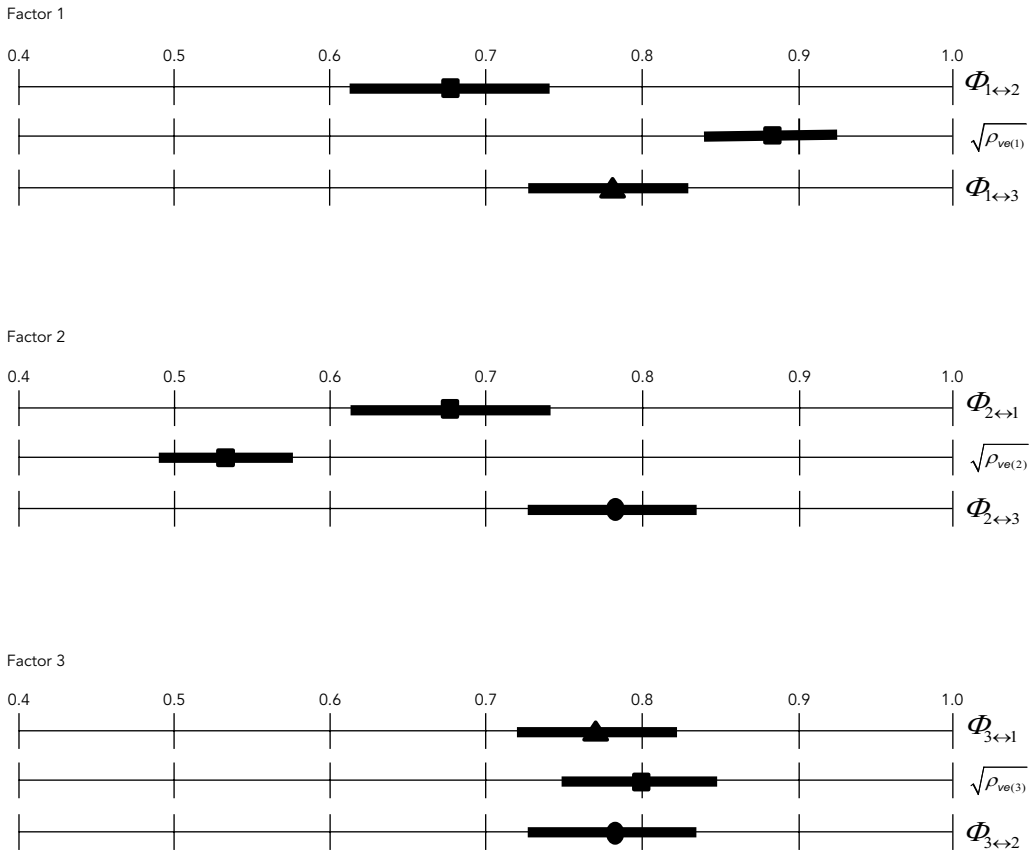
Figure 1 portrays a hypothetical scenario involving a three-factor structure and different strengths of FDV. While $\sqrt{\overline{P}_{ve(1)}}$ for Factor 1 is plainly higher than both its correlations with Factors 2 ($\Phi_{1\leftrightarrow2}$) and 3 ($\Phi_{1\leftrightarrow3}$) – i.e., FDV seems corroborated –, the scenario concerning Factor 2 shows quite the opposite, with $\sqrt{\overline{P}_{ve(2)}}$ strikingly below the respective factor correlations ($\Phi_{2\leftrightarrow1}$ and $\Phi_{2\leftrightarrow3}$). FDV would not hold here. The situation regarding Factor 3 is less clear since the overlaps of all three 95% confidence intervals are far from inconsequential. The differences between $\sqrt{\overline{P}_{ve(3)}}$ and both factor correlations ($\Phi_{3\leftrightarrow1}$ and $\Phi_{3\leftrightarrow2}$) require formal testing before any decision is made. Note that, given the estimates depicted in Figure 1, a researcher could easily be misled by following a commonly held rule-of-thumb used in applied research that only regards a factor correlation ≥ 0.85 as offering evidence for poor discriminant validity [9,25,26].

The absence of discriminant factorial validity may be the result of poor dimensional structure specification, meaning, for instance, that the two highly correlated factors supposedly covering different dimensions of a construct form a one-dimensional rather than the conjectured two-dimensional structure. An exploratory approach may be used to investigate this hypothesis.

Sometimes, though, there remains a signal from the data that separate factors do exist, albeit the highly correlated factors. In this case, a higher-order factorial model may be considered. Fitting and statistically testing these models requires more than two factors and a minimum of component items *per* factor [9]. An alternative consists of exploring a general factor, which as-

Figure 1

Example of a scenario involving three-factor structure and different degrees of factor-based discriminant validity.



Φx↔y: factor correlations; *Pve( ):* average variance extracted of factor *x* (in brackets: 95% confidence intervals).

sumes that the complete set of component items prominently load on a single all-encompassing factor, along with, or in spite of, the originally postulated specific factors. These are called bi-factor models [18,27,28]. Although proposed over half a century ago [29], bi-factor models have recently gained renewed interests and software developments involving bi-factor EFA, ESEM and Bayesian models [18].

Another possibility is that there are unspecified cross-loadings unduly attempting to "express themselves" through what could be thought of as a "backdoor" path, i.e., by circulating information (signal) through pumping up factor correlations. The solution is clearly to identify first these cross-loadings, and thereafter recalculate and assess FDV. Of course, the uncovered cross loadings would still require attention as discussed in a previous section (Step 2).

In closing this section, a word is due on how factorial convergent/discriminant validity intertwines with internal consistency. The latter is a property frequently reported related to the notion of reliability and though traditionally estimated through intra-class correlation coefficient [30], it may be recast in terms of the factor-based estimates [10,31] (cf. Webappendix, note 6; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf).

Step 5: evaluating item discrimination and intensity *vis-à-vis* the latent trait spectrum

Many epidemiological measurement tools hold categorical items and in these cases it is also useful to evaluate their ability to discriminate subjects along the latent trait *continuum*. For this purpose, we may turn to Item Response Theory (IRT) models [14,32]. Also known as latent trait theory, IRT allows relating the characteristics of items and subjects to the probability of endorsing a particular response category. IRT models are commended when the latent variable is assumed continuous and used to explain the response of the individual to dichotomous or polychotomous indicators [33]. A review of IRT is beyond the scope of this text, but for a better understanding of what ensues, the reader may want to consult the Webappendix, note 7, for a brief account on the underlying assumptions [7]; alternative IRT models [34,35]; and the related discrimination (a_i) and intensity (b_i) parameters. Further details may be found in Streiner & Norman [36], van der Linden & Hambleton [37], Embretson & Reise [32], De Boeck & Wilson [38], Ayala [39], Hardouin [40], and many references therein.

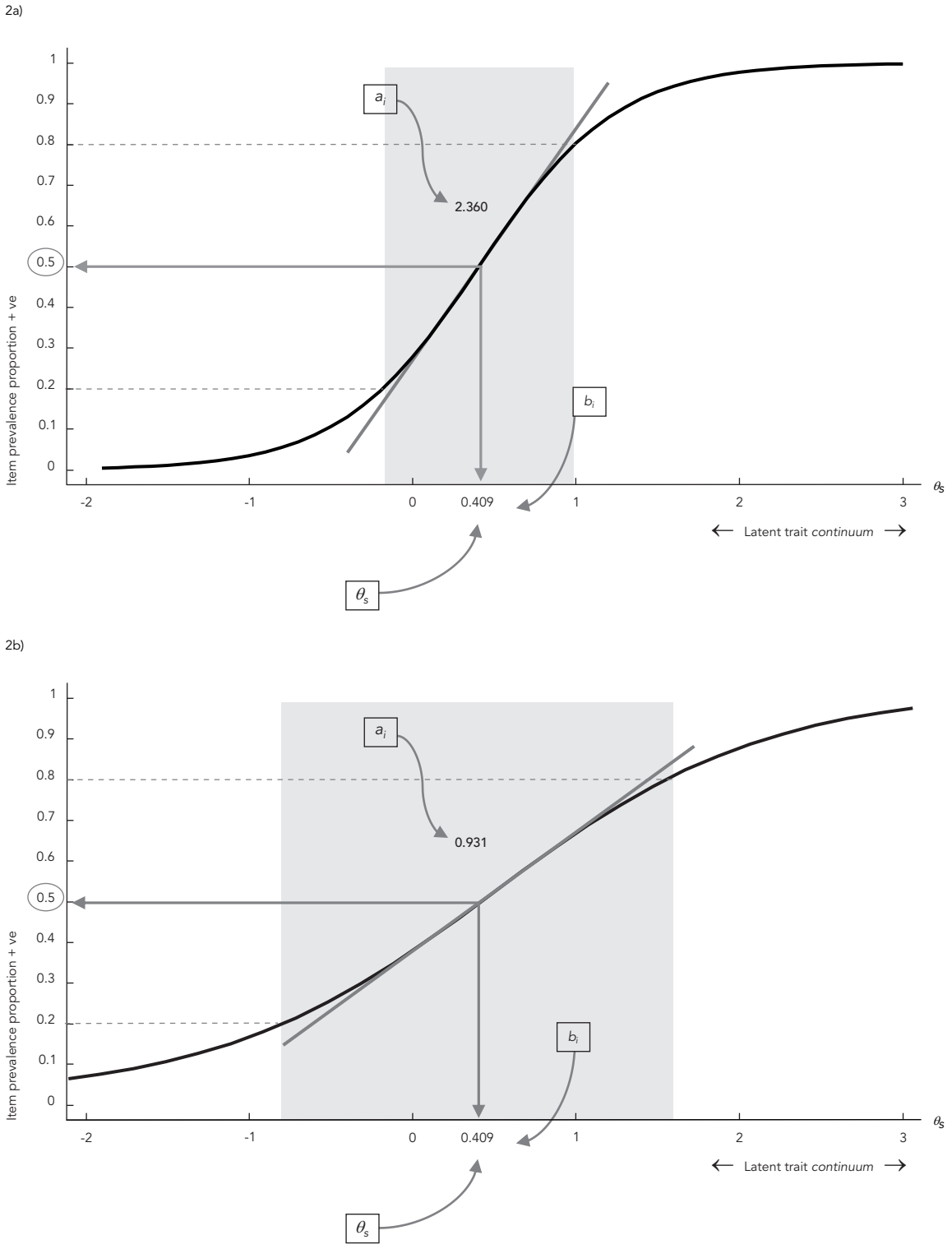
Within the context of an instrument's dimensional (internal) evaluation, IRT models may be regarded as one-dimensional nonlinear factor models of a confirmatory type [9,11,41,42]. If a CFA model is parameterized in a certain way – by freely estimating all loadings and thresholds and setting factor variances to 1 – there is a direct relationship between obtained factor loadings λ_i and the IRT a_i parameters of interest in this section. The direct relationship is given by $a_i = \lambda_i / \sqrt{1-\lambda_i^2}$, indicating that the discrimination parameter is simply the ratio of the item's loading to its residual variance (*uniqueness*) since $\delta_1 = 1-\lambda_i^2$. This ratio is thus the amount of information the item shares with the latent trait to what it does not [9,17].

Larger values of a_i correspond to steeper item characteristic curves (ICC), indicating that if an item has good discriminant ability vis-à-vis the level of the construct where it is located along the spectrum, for any given level of the latent trait θ, there is a rapid change in probability of response in the positive direction. Conversely, small values of a_i correspond to less inclined curves showing that the positive response probability increases rather slowly and that subjects with low levels of the latent trait may have similar probabilities of endorsing a given item than subjects at higher levels of the continuum. The corresponding item response curves are respectively illustrated in Figure 2a and 2b. Note the difference in "coverage" between the two ICC. The probability of a positive response in this most discriminating scenario (a) increases from 0.2 to 0.8 within a quite narrow spectrum of values of the latent trait – θ_s roughly varying from -0.10 to +0.97 – while in scenario (b) the probability increases at a much slower rate, now covering a wider range of θ_s – roughly, -0.70 to +1.5.

The intensity parameter b_i represents the location of response categories along the continuum of the latent variable. These parameters may also be estimated through a specifically parameterized CFA model since they are a function of the obtained k–1 thresholds (τ_i) of categorical items with k levels. This relation is given by $b_i = \tau_i / \lambda_i$, where λ_i are the i factor loadings. Note that if items are dichotomous there will be only one threshold per item and thus only one *b*-parameter per item. This parameter corresponds to the latent trait level wherein there is a 50% chance of change in response category (e.g., from negative to positive), conditional on the subject's propensity level along the latent trait (Figure 2a and 2b again). In Samejima's graded response model [35], for each item, there are as many ICC as there are k – 1 cut-off points between categories. In a "well-behaved" instrument, one thus expects to have increasing values of b_{ik} (where the subscript now indicates

Figure 2

Two examples of item characteristic curves.



a_i: item discrimination parameter; b_i: item difficulty parameter; θ_s: latent trait level.

threshold $k \geq 2$ of item i), while also a gradient across items filling up the θ_s spectrum.

From an interpretative viewpoint, as a complement to what has been exposed in Step 3 regarding content redundancy via measurement error correlation, one could think of a second type of "redundancy" when two or more items have overlapping b_i location parameters. Although not necessarily sharing meaning and content, they would still be mapping similar regions of the latent trait. Too many overlaying indicators may lead to inefficiency since a "latent intensity region" would end up being superfluously tapped. The result would be that much interviewing time would be unduly spent in repeatedly asking similarly functioning questions (indicators), but with effectively little discretion in regards to the desired rise in "latent intensity".

Thus, accepting a set of items as effective "mappers" of a latent trait goes beyond the mere sustainability of the purported dimensional structure and the reasonableness of the respective loading magnitudes. It also depends on how the item thresholds span along the latent spectrum. Yet, information clustering and ensuing "coverage redundancy" is only one side of the problem requiring inspection. The other concerns the information gaps that may possibly be left open along the latent trait region. It may well happen that some b_i parameters end up clustering in a region depicting a lower intensity of the θ_s latent spectrum, whereas other items group on the opposite "more intense" side. This void leading to sparse information in between would be clearly undesirable. Although an overall score would still be possible (be it in a factor-based or a raw metric), mapping the construct would not be smooth, entailing information gaps in some regions of the *continuum*, along with information overlapping in others. Further scrutiny through misfit diagnostics would be welcome so that decisions to modify or even remove items from a measurement tool are evidence-based rather than anchored on mere appearance [40,43]. Face validity as to which component items are to be included, modified or substituted may be an important starting point, but any changes need sound support.

Step 6: examining raw scores as latent factors score proxies

Although, in practice, model-based factor scores may be desirable and are estimable – either implicitly whilst estimating causal parameters in complex structural equation models, or explicitly by drawing plausible latent value from the Bayesian posterior distribution [19] –, in many applied epidemiological studies it is common to use raw scores as their empirical representations. A scale's total raw score is typically calculated by summing up the component items' raw scores and used "as is" to rank respondents along the *continuum* of the latent variable, or sometimes after categorization following pre-defined cut-off points. Regardless, before a total raw score can be used in practice, it is essential to verify how it relates to the corresponding factor score – the most plausible representative of the latent variable – and to have its psychometric properties scrutinized.

This evaluation may start with examining the correlation between the raw score and the model-based factor score. Once a strong correlation is established, and so implying that the raw score closely charts the factor score, scalability and monotonicity should be sought. This may be carried out via non-parametric item-response theory (NIRT) methods [7].

Scalability refers to the ability of items and, by extension, the overall ensuing score of a scale to order and properly position subjects along the *continuum* of the latent trait. Besides items covering evenly the entire spectrum of the latent variable (Step 5), it is also expected that these items and, by extension, the overall composite score are able to seize an ascending gradation of intensity. The underlying assumption is that if there are items with increasing intensity on a scale, a subject scoring positively on a given i^{th} *i*tem will have already scored positively on all items of lesser intensity. This scenario would constitute the perfect Guttman scalogram [8] (cf. Webappendix, note 8; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf). Since this ideal pattern materializes seldom (if ever) in real data, the key is to test whether such an underlying state can be assumed as giving rise to the actual data at hand. Scalability may be gauged through Loevinger's H coefficient [7] (cf. Webappendix, note 9; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf). As suggested by Mokken, values > 0.3 indicate that the scalability assumption is acceptable, whereas values close to 1.0 indicate that the items form a near perfect Guttman scalogram [7].

Under the Monotone Homogeneity Model (MHM), the monotonicity assumption holds when the probability of an item response greater than or equal to any fixed value is a nondecreasing function of the latent trait θ_s [44]. For scales involving dichotomous items, this satisfies by showing scalability. Unlike the MHM, however, a Double Monotonicity Model also assumes that the Item Response Functions (IRF) do not intersect across items (*a.k.a.* invariant item ordering).

For scales formed by polychotomous items, the $k \geq 2$ Item Step Response Functions (ISRF) of any given item containing $k + 1$ levels may not intersect if the monotonicity assumption is sustained. When the double monotonicity assumption also holds, besides "within item" monotonicity (and ensuing nonintersections of the k ISRFs), nonintersections should also occur across ISRFs of different items [7]. Under double monotonicity, one may be fairly confident that items' scores are answered and thus interpreted consistently by all respondents, whatever their level of the latent trait [7].

Single and double monotonicity may be evaluated through the criteria proposed by Molenaar et al. [45]. Accordingly, a criterion less than 40 suggests that the reported violations (response function intersections) may be ascribed to sampling variation. If the criterion is between 40 and 80, a more detailed evaluation is warranted. A criterion beyond 80 raises doubts about the monotonicity assumption of an item and in turn, about the scale as a whole. Additionally, assessing the number and percentage of violations of monotonicity may also help in the examination. Monotonicity may also be inspected graphically by means of the item traces as a function of the "rest score" formed by the raw score in which the item in focus is left out. See Reichenheim et al. [46] for an applied example with display. A full account of the methods employed here and details on NIRT may be found in Molenaar et al. [45], Sijtsma & Molenaar [7], and Hardouin et al. [44].

Step 7: assessing dimensional structure and measurement invariance across groups

Ideally, the psychometric properties of an instrument and hence its overall performance should be stable across different population groups (e.g., gender, age, occupations, regions, cultures). Unsupported measurement invariance suggests problems in the design of the instrument, which might compromise inferences and comparisons between groups.

Invariance assessment can be accomplished by multiple-group confirmatory factor analysis (MG-CFA), MIMIC models (*Multiple Indicators, Multiple Causes* models, a.k.a., AFC with covariates) and IRT models [9]. Although respective model assumptions and test procedures differ, each approach may be regarded as a particular case of the generalized latent variable modelling framework [33,47,48].

In MG-CFA models, a measurement model is estimated simultaneously in several subgroups. These models offer the advantage that equivalence of all parameters described above may be gauged at once and making possible the simultaneous evaluation of dimensional and measurement invariance. The approach consists of testing a set of measurement models in a systematic sequence, for instance (adapted from Kankaraš et al. [33] and Milfont & Fischer [49]), by (1) specifying a factorial model for each sample (group); (2) evaluating samples simultaneously to determine whether the factor structure is identical when all parameters are freely estimated (configural invariance); (3) testing loading invariance by freely estimating loadings in one group and constraining all values of the second group to a symmetrical equality (metric invariance); and (4) additionally examining intercept/threshold equality across groups (scalar invariance) (cf. Webappendix, note 10; http://www.ims.uerj.br/docentes/documentos/39/CSP_1436_13_appendix.pdf).

The appraisal of equivalence is achieved in comparing the parameter estimates and fit indices of the model. Besides visually inspecting estimated parameters and evaluating *per* group fit indices, change in the fit of nested models should also be assessed by, e.g., contrasting the chi-square of a model with all parameters freely estimated in both groups with another in which the corresponding item parameters are constrained to be the same. A non-significant chi-square difference would favor equivalence whereas the absence of equality in at least one parameter would support the rejection of the null hypothesis.

Although this approach seems straightforward in principle, the overall assessment of invariance – the anticipated "universalist" scenario – may become quite unmanageable as the dimensional structure becomes more complex. Considering all loadings/residuals and thresholds tested in a multi-dimensional system, the prospect of ending up identifying an invariance violation becomes real. Moreover, the rejection of estimate differences across groups also depends on the sample size. Although some recommended goodness-of-fit indices such as RMSEA or CFI account for sample size and model complexity, likelihood ratio chi-squares are typically used to assess nested models, and it is quite likely that minor differences in estimates across groups will flag statistical significances. The point to make is whether, for instance, a "statistically different" loading of $\lambda_{1(G1)} = 0.85$ in one group should actually be considered "substantively different" from a loading of $\lambda_{1(G2)} = 0.77$ in another group. Recent statistical developments using ML-based and Bayesian multi-group factorial models are promising since they allow relaxing the strict zero null hypothesis [18,50,51].

A MIMIC model consists of a regression of group indicators (e.g., gender: male = 0 and 1 =

female) on latent factors and sometimes on indicators (items). Since they are much simpler to specify and only allow estimating item intercepts or threshold invariances, these models could perhaps be thought of as a preliminary testing stage.

IRT models also allow assessing parameter invariance across groups. Both slope (a_i) and intercept/threshold (b_i) invariances may be inspected. The latter is most often reported [32] and referred to as uniform differential item functioning (DIF) in the literature. Departure from slope invariances are designated nonuniform DIF [33]. Unlike MG-CFA, the IRT approach starts from the most restricted model, in which all parameters are constrained to equality across groups. This baseline model is then compared with models in which item parameters are allowed to vary freely across groups, one at a time [52]. An IRT approach may be advantageous is some circumstances such as when items are ordinal since it assesses several b_{ik} parameters per item for each $k \geq 2$ threshold, whereas only one threshold can be estimated for each item in a traditional MG-CFA. However, IRT requires a one-dimensional construct, larger sample sizes and more items per scale for statistical efficiency, and works better when invariance is evaluated in no more than two groups [52].

Discussion

Although arranged sequentially for didactic purposes, the seven steps presented in this article constitute an iterative process in practice. In the process of developing a new instrument, for instance, a large factor correlation detected in Step 4 may raise suspicions of a hidden configural misspecification, commending Step 1 to be re-visited. As another example, two quite "well behaved" factors may, in fact, hold an underlying effects method [22]. In these cases, it would be more appropriate to take a step back and, guided by theory, revise the entire measurement model.

From a systematic review standpoint, at least inspecting all steps would be highly recommended. It may well be that the evidence concerning an instrument under scrutiny is scattered and incomplete, but still, holding this scrutiny against some intelligible standard may help in identifying gaps and pointing out future research. This reminds us that the development of any given instrument involves a laborious and long process – including replications for consistency – and that a decision to promote and endorse its use cannot lean on only a few restricted and frail explorations. A lot of maturing is required before a "quality tag" may be assigned to an instrument.

Several issues would still be worth exploring and could perhaps be added to the proposed roadmap for assessing the quality of studies on the structural validity of epidemiological measurement instruments. One is the analysis of invariance through Bayesian models, which allow, for instance, relaxing the constraint of setting cross-loadings to absolute zeros in CFA-type models [18]. Another issue that requires refinement concerns the process of identifying cut-off points on scales composed of raw scores that is based on a covariance modelling (e.g., latent class analysis [17,53]), rather than relying on some untested *a priori* rationale or worse, simply by arbitrarily setting thresholds at fixed and equidistant intervals. This appraisal could even qualify as another full evaluation step since many epidemiological tools are often recommended and used as categorical variables.

The indefinite article adopted in the title – "*a* seven-step roadmap" – purposely conveys the notion that the current proposal is clearly incomplete and still a process under construction, certainly requiring refinements to make it more comprehensive and operational to the final user. It should be seen as an attempt to synthesize information in an effort to add substance to the COSMIN initiative in promoting a common and systematic approach aimed at granting robust "quality labels" to measurement tools used in health research. Perhaps, in the long run, clear and practical guidelines may ensue from the discussions initiated here.

Resumen

*Se han propuesto directrices para evaluar la calidad de los ensayos clínicos, estudios observacionales y estudios de validación de pruebas de diagnóstico. Más recientemente, la iniciativa COSMIN (*COnsensus-based Standards for the selection of Health Measurement INstruments*) amplió estas directrices para la medición de herramientas epidemiológicas en general. Una de las muchas facetas propuestas para la evaluación se refiere a la validez de la estructura dimensional del instrumento (o validez estructural). El objetivo de este trabajo es extender estas directrices. Se propone un guion de siete pasos, examinando: (1) la estructura dimensional, (2) la fuerza de indicadores componentes relativos con los patrones de las cargas y errores de medición; (3) la correlación de los residuos, (4) la validez factorial convergente y discriminante; (5) la capacidad de discriminación e intensidad de indicadores en relación a rasgos latentes; (6) las propiedades de las puntuaciones brutas, (7) invariancia factorial. El artículo también señala que las medidas propuestas aún requieren mayor discusión y están abiertos a mejoras.*

Modelos Epidemiológicos; Validez de las Pruebas; Metodología

Contributors

M. E. Reichenheim contributed in planning and writing up all drafts and the final version of the manuscript. Y. H. M. Hökerberg contributed in planning and writing up all drafts and the final version of the manuscript. C. L. Moraes contributed to writing the final version of the manuscript.

Acknowledgments

M.E.R. was partially supported by the CNPq (process n. 301221/2009-0). C.L.M. was partially supported by the CNPq (process n. 302851/2008-9) and Faperj (process n. E-26/110.756/2010).

References

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138:W1-12.
2. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003; 3:25.
3. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996; 18: 979-92.
4. McDowell I, Jenkinson C. Development standards for health measures. J Health Serv Res Policy 1996; 1:238-46.
5. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002; 11:193-205.
6. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010; 19:539-49.
7. Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. Thousand Oaks: Sage Publications; 2002.
8. Wilson M. Constructing measures. An item response modeling approach. Mahwah: Lawrence Erlbaum Associates; 2005.
9. Brown TA. Confirmatory factor analysis for applied research. New York: Guilford Press; 2006.
10. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. SEM: confirmatory factor analysis. In: Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL, editors. Multivariate data analysis. Upper Saddle River: Pearson Prentice Hall; 2006. p. 770-842.
11. Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Boca Raton: Chapman & Hall/CRC; 2004.
12. Gorsuch RL. Factor analysis. Hillsdale: Lawrence Erlbaum; 1983.
13. Gerbing DW, Hamilton JG. Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. Struct Equ Modeling 1996; 3:62-72.
14. Hancock GR, Mueller RO. The reviewer's guide to quantitative methods in the social sciences. New York: Routledge; 2010.
15. Jöreskog KG. Testing structural equation models. In: Bollen KA, Long JS, editors. Testing structural equation models. London: Sage Publications; 1993. p. 294-316.

16. Marsh HW, Muthén B, Asparouhov A, Lüdtke O, Robitzsch A, Morin AJS, et al. Exploratory structural equation modeling, integrating CFA and EFA: application to students' evaluations of university teaching. Struct Equ Modeling 2009; 16:439-76.
17. Muthén B, Asparouhov T. Latent variable analysis with categorical outcomes: multiple-group and growth modeling in Mplus. Mplus Web Notes 2002; (4). https://www.statmodel.com/examples/webnote.shtml#web4.
18. Muthen B, Asparouhov T. Bayesian structural equation modeling: a more flexible representation of substantive theory. Psychol Methods 2012; 17:313-35.
19. Muthén LK, Muthén BO. Mplus user's guide. 7th Ed. Los Angeles: Muthén & Muthén; 1998/2012.
20. Kline RB. Principles and practice of structural equation modeling. New York: Guilford Press; 2011.
21. Byrne BM. Structural equation modeling with Mplus: basic concepts, applications, and programming. New York: Routledge; 2012.
22. Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003; 88:879-903.
23. Marsh HW. Positive and negative global self-esteem: a substantively meaningful distinction or artifactors? J Pers Soc Psychol 1996; 70:810-9.
24. Fornell C, Larcker DF. Evaluating structural equation models with unobservable variables and measurement error. J Market Res 1981; 18:39-50.
25. Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. Mahwah: Lawrence Erlbaum Associates; 2003.
26. Tabachnick BG, Fidell LS. Using multivariate statistics. Boston: Allyn & Bacon; 2001.
27. Cai L. A two-tier full-information item factor analysis model with applications. Psychometrika 2010; 75:581-612.
28. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Qual Life Res 2007; 16 Suppl 1:19-31.
29. Holzinger KJ, Swineford F. A study in factor analysis: the stability of a bi-factor solution. Chicago: University of Chicago Press; 1939.
30. Henson RK. Understanding internal consistency reliability estimates: a conceptual primer on coefficient alpha. Meas Eval Couns Develop 2001; 34:177-89.
31. Raykov T, Shrout P. Reliability of scales with general structure: point and interval estimation using a structural equation modeling approach. Struct Equ Modeling 2002; 9:195-212.
32. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah: Lawrence Erlbaum Associates; 2000.
33. Kankaraš M, Vermunt JK, Moors G. Measurement equivalence of ordinal items: a comparison of factor analytic, item response theory and latent class approaches. Sociol Meth Res 2011; 40:279-310.
34. Rasch G. Probablistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
35. Samejima F. Graded response model. In: van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer-Verlag; 1996. p. 85-100.
36. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Oxford: Oxford University Press; 2008.
37. van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer-Verlag; 1996.
38. De Boeck P, Wilson M. Explanatory item response models: a generalized linear and nonlinear approach. New York: Springer-Verlag; 2004.
39. Ayala RJ. The theory and practice of item response theory. New York: Guilford Press; 2009.
40. Hardouin J-B. Rasch analysis: estimation and tests with raschtest. Stata J 2007; 7:22-44.
41. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychol Bull 1993; 114:552-66.
42. Takane Y, de Leeuw J. On the relationship between item response theory and factor analysis of discretized variables. Psychometrika 1987; 52: 393-408.
43. Smith RM, Suh KK. Rasch fit statistics as a test of the invariance of item parameter estimates. J Applied Meas 2002; 4:153-63.
44. Hardouin JB, Bonnaud-Antignac A, Sebille V. Nonparametric item response theory using Stata. Stata J 2011; 11:30-51.
45. Molenaar IW, Sijtsma K, Boer P. MSP5 for Windows. User's manual for MSP5 for Windows: a program for mokken scale analysis for polytomous items (version 5.0). Groningen: iec ProGAMMA; 2000.
46. Reichenheim ME, Moraes CL, Oliveira ASD, Lobato G. Revisiting the dimensional structure of the Edinburgh Postnatal Depression Scale (EPDS): empirical evidence for a general factor. BMC Med Res Methodol 2011; 11:94.
47. Rabe-Hesketh S, Skrondal A. Classical latent variable models for medical research. Stat Meth Med Res 2008; 17:5-32.
48. Skrondal A, Rabe-Hesketh S. Latent variable modelling: a survey. Scand J Stat 2007; 34.
49. Milfont TL, Fischer R. Testing measurement invariance across groups: applications in cross-cultural research. Int J Psychol Res 2010; 3:111-30.
50. Muthén B, Asparouhov T. BSEM measurement invariance analysis. Mplus Web Note 2013; (17). https://www.statmodel.com/examples/webnotes/webnote18.pdf.
51. Asparouhov T, Muthén B. Multiple-group factor analysis alignment. Struct Equ Modeling; in press.
52. Meade AW, Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. Organizational Research Methods 2004; 7:361-88.
53. Geiser C. Data analysis with Mplus. New York: Guilford Press; 2003.

Submitted on 07/Aug/2013
Final version resubmitted on 25/Nov/2013
Approved on 17/Dec/2013