Psicothema

Psicothema 2014, Vol. 26, No. 1, 108-116
doi: 10.7334/psicothema2013.260

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Validity evidence based on internal structure

Joseph Rios and Craig Wells
University of Massachusetts Amherst (USA)

Abstract

Background: Validity evidence based on the internal structure of an assessment is one of the five forms of validity evidence stipulated in the Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. In this paper, we describe the concepts underlying internal structure and the statistical methods for gathering and analyzing internal structure. **Method:** An in-depth description of the traditional and modern techniques for evaluating the internal structure of an assessment. **Results:** Validity evidence based on the internal structure of an assessment is necessary for building a validity argument to support the use of a test for a particular purpose. **Conclusions:** The methods described in this paper provide practitioners with a variety of tools for assessing dimensionality, measurement invariance and reliability for an educational test or other types of assessment.

Keywords: validity, standards, dimensionality, measurement invariance, reliability.

Resumen

Evidencia de validez basada en la estructura interna. **Antecedentes:** la evidencia de validez basada en la estructura interna de una evaluación es una de las cinco formas de evidencias de validez estipuladas en los *Standards for Educational and Psychological Testing* de la *American Educational Research Association, American Psychological Association, and National Council on Measurement in Education*. En este artículo describimos los conceptos que subyacen a la estructura interna y los métodos estadísticos para analizarla. **Método:** una descripción detallada de las técnicas tradicionales y modernas para evaluar la estructura interna de una evaluación. **Resultados:** la evidencia de validez basada en la estructura interna de una evaluación es necesaria para elaborar un argumento de validez que apoye el uso de un test para un objetivo particular. **Conclusiones:** los métodos descritos en este artículo aportan a los profesionales una variedad de herramientas para evaluar la dimensionalidad, invarianza de la medida y fiabilidad de un test educativo u otro tipo de evaluación.

Palabras clave: validez, standards, estructura interna, dimensionalidad, invarianza de la medida, fiabilidad.

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) list five sources of evidence to support the interpretations and proposed uses of test scores: evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing. According to the *Standards*, evidence based on internal structure, which is the focus of this paper, pertains to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13).

There are three basic aspects of internal structure: dimensionality, measurement invariance, and reliability. When assessing dimensionality, a researcher is mainly interested in determining if the inter-relationships among the items support the intended test scores that will be used to draw inferences. For example, a test that intends to report one composite score should be predominately unidimensional. For measurement invariance, it is useful to provide evidence that the item characteristics (e.g., item discrimination and difficulty) are comparable across manifest groups such as sex or race. Lastly, reliability indices provide evidence that the reported test scores are consistent across repeated test administrations. The purpose of the present paper is to describe basic methods for providing evidence to support the internal structure of a test (e.g., achievement tests, educational surveys, psychological inventories, or behavioral ratings) with respect to assessing dimensionality, measurement invariance, and reliability.

Received: August 27, 2013 • Accepted: October 10, 2013
Corresponding author: Joseph Rios
School of Education
University of Massachusetts Amherst
01003 Amherst (Estados Unidos)
e-mail: jarios@educ.umass.edu

Assessing dimensionality

Assessing test dimensionality is one aspect of validating the internal structure of a test. Factor analysis is a common statistical method used to assess the dimensionality of a set of data (Bollen, 1989; Brown, 2006; Kline, 2010; Thompson, 2004). There are several factor analytic methods available for analyzing test dimensionality; however, this paper will focus solely on confirmatory factor analysis, which is the most comprehensive means for comparing hypothesized and observed test structures.

Confirmatory factor analysis

Confirmatory factor analysis (CFA) is a type of structural equation model (SEM) that examines the hypothesized

relationships between indicators (e.g., item responses, behavioral ratings) and the latent variables that the indicators are intended to measure (Bollen, 1989; Brown, 2006; Kline, 2010). The latent variables represent the theoretical construct in which evidence is collected to support a substantive interpretation. In comparison to exploratory factor analysis (EFA), a basic feature of CFA is that the models are specified by the researcher a priori using theory and often previous empirical research. Therefore, the researcher must explicitly specify the number of underlying latent variables (also referred to as factors) and which indicators load on the specific latent variables. Beyond the attractive feature of being theoretically driven, CFA has several advantages over EFA such as its ability to evaluate method effects and examine measurement invariance.

CFA provides evidence to support the validity of an internal structure of a measurement instrument by verifying the number of underlying dimensions and the pattern of item-to-factor relationships (i.e., factor loadings). For example, if the hypothesized structure is not correct, the CFA model will provide poor fit to the data because the observed inter-correlations among the indictors will not be accurately reproduced from the model parameter estimates. In this same vein, CFA provides evidence of how an instrument should be scored. If a CFA model with only one latent variable fits the data well, then that supports the use of a single composite score. In addition, if the latent structure consists of multiple latent variables, each latent variable may be considered a subscale and the pattern of factor loadings indicates how the subscores should be created.

If the multi-factor model fits the data well, and the construct is intended to be multidimensional, then that is evidence supporting the internal structure of the measurement instrument. Furthermore, for multi-factor models, it is possible to assess the convergent and discriminant validity of theoretical constructs. Convergent validity is supported when indicators have a strong relationship to the respective underlying latent variable. Discriminant validity is supported when the relationship between distinct latent variables is small to moderate. In fact, CFA can be used to analyze multitrait-multimethod (MTMM; Campbell & Fisk, 1959) data (Kenny, 1976; Marsh, 1989).

Three sets of parameters are estimated in a CFA model. For one, the factor loadings, which represent the strength of the relationship between the indicator and its respective latent variable and may be considered a measure of item discrimination, are estimated. In CFA, the factor loadings are fixed to zero for indicators that are not hypothesized to measure a specific latent variable. When standardized, and no cross-loadings exist (i.e., each indicator loads on one latent variable), the factor loadings may be interpreted as correlation coefficients. The variance and covariance coefficients for the latent variables are also estimated. However, the variance for each latent variable is often fixed to one to establish the scale of the latent variable. Fixing the variance for each latent variable to one produces a standardized solution. Lastly, the variance and covariance coefficients for the measurement errors (i.e., unique variance for each indicator) are estimated. When the measurement errors are expected to be uncorrelated, the covariance coefficients are fixed to zero.

To examine the internal structure of a measurement instrument, the CFA model is evaluated for model fit and the magnitude of the factor loadings and correlations among the latent variables are examined. Model fit determines if the hypothesized model can reproduce the observed covariance matrix (i.e., covariance matrix for the indicators) using the model parameter estimates. If the model is specified incorrectly (e.g., some indicators load on other latent variables) then the model will not fit the data well. Although there are several approaches to assess model fit, such as hypothesis testing, the most common method uses goodness-of-fit indices. There are a plethora of goodness-of-fit indices available for a researcher to use to judge model fit (see Bollen, 1989; Hu & Bentler, 1999). It is advisable to use a few of the indices in evaluating model fit. Some of the more commonly used indices are the comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Suggested cutoff values are available to help researchers determine if the model provides adequate fit to the data (e.g., See Hu & Bentler, 1999). A model that does not fit the data well must be re-specified before interpreting the parameter estimates. Although there are numerous CFA models that one can fit to the sample data, in this paper we describe and illustrate the increasingly popular bifactor model.

Bifactor model

The bifactor model (also referred to as the *nested* or *general-specific* model) first introduced by Holzinger and Swineford (1937) has seen a drastic increase in popularity within the SEM and item response theory (IRT) literature over the past few years. Once overshadowed by alternative multidimensional models, such as the correlated-factors and second-order models, advances in parameter estimation, user-friendly software, and novel applications (e.g., modeling differential item functioning (Fukuhara & Kamata, 2011; Jeon, Rijmen, & Rabe-Hesketh, 2013), identifying local dependence (Liu & Thissen, 2012), evaluating construct shift in vertical scaling (Li & Lissitz, 2012), to name a few) have led to a renewed interest in the model. However, applications of the bifactor model have been limited in the field of psychology, which some have suggested is due to a lack of familiarity with the model and a lack of appreciation of the advantages it provides (Reise, 2012). Therefore, the objective of the current section is to provide a general description of the confirmatory canonical bifactor model, note some of the advantages and limitations associated with the model, and discuss techniques for determining model selection when comparing unidimensional and bifactor models.

General description of bifactor model. The bifactor model is a multidimensional model that represents the hypothesis that several constructs, as indicated each by a subset of indicators, account for unique variance above and beyond the variance accounted for by one common construct that is specified by all indicators. More specifically, this model is composed of one *general* and multiple *specific* factors. The general factor can be conceptualized as the target construct a measure was originally developed to assess, and accounts for the common variance among all indicators. In contrast, specific factors pertain to only a subset of indicators that are highly related in some way (e.g., content subdomain, item type, locally dependent items, etc.), and account for the unique variance among a subset of indicators above and beyond the variance accounted for by the general factor. Within the confirmatory model, each indicator loads on the general factor *and* on one and only one specific factor. Allowing indicators to cross-load on multiple specific factors leads to questionable parameter estimates, and is limited by the small degrees of freedom available in the model. As

the specific factors are interpreted as the variance accounted for above and beyond the general factor, an orthogonal (uncorrelated) assumption is made for the relationships between the general and specific factors. Furthermore, the covariances among the specific factors are set to 0 to avoid identification problems (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012). The residual variances of the indicators are interpreted as the variance unaccounted for by either the general or specific factors (see Figure 1). Within the field of psychology, this model has been applied to study a number of constructs, such as depression (Xie et al., 2012), personality (Thomas, 2012), ADHD (Martel, Roberts, Gremillion, von Eye, & Nigg, 2011), and posttraumatic stress disorder (Wolf, Miller, & Brown, 2011).

Advantages of the bifactor model. The bifactor model possesses the following four advantages over other multidimensional models (e.g., the second-order model): 1) the domain specific factors can be studied independently from the general factor, 2) the relationship between the specific factors and their respective indicators can be evaluated, 3) invariance can be evaluated for both the specific and general factors independently, and 4) relationships between the specific factors and an external criterion can be assessed above and beyond the general factor (Chen, West, & Sousa, 2006). The ability to study the specific factors independently from the general factor is important in better understanding theoretical claims. For example, if a proposed specific factor did not account for a substantial amount of variance above and beyond the general factor, one would observe small and non-significant factor loadings on the specific factor, as well as a non-significant variance of the specific factor in the bifactor model. This would notify the researcher that the hypothesized specific factor does not provide unique variance beyond the general factor, which would call for a modification of the theory and the test specifications. A closely related advantage of the bifactor model is the ability to directly examine the strength of the relationship between the specific factors and their respective indicators. Such an assessment provides a researcher with information regarding the appropriateness of using particular items as indicators of the specific factors. If a relationship is weak, one can conclude that the item may be appropriate solely as an indicator of the general factor.

The last two advantages deal directly with gathering validity evidence to support a theoretical rationale. More specifically, within the bifactor model one has the ability to evaluate invariance for both the specific and general factors independently. This would allow researchers to directly compare means of the latent factors (both the specific and general factors), if scalar invariance is met, across distinctive subgroups of examinees within the population (See Levant, Hall, & Rankin, 2013). Lastly, the bifactor model is advantageous in that one can study the relationships between the specific factors and an external criterion or criteria above and beyond the general factor. This application of the bifactor model could be particularly attractive for gathering evidence based on relations to other variables (convergent and discriminant evidence, as well as test-criterion relationships) for multidimensional measures.

Limitations of the bifactor model. Although the bifactor model provides numerous advantages, it is also has some limitations. As noted by Reise, Moore, and Haviland (2010), there are three major reasons for limiting the application of the bifactor model in practice: 1) interpretation, 2) model specification, and 3) restrictions. The first major limiting factor for practitioners is relating the bifactor model to their respective substantive theories. More specifically, the bifactor model assumes that the general and specific factors are orthogonal to one another, which may be too restrictive or make little sense in adequately representing a theoretical model. For example, if one were studying the role of various working memory components on reading skills, it would be difficult to assume the relationship between these two constructs is orthogonal. Instead, competing multidimensional models, such as the correlated-traits or second-order models would be more attractive as the restrictive orthogonality assumption is not required. This is one of the major

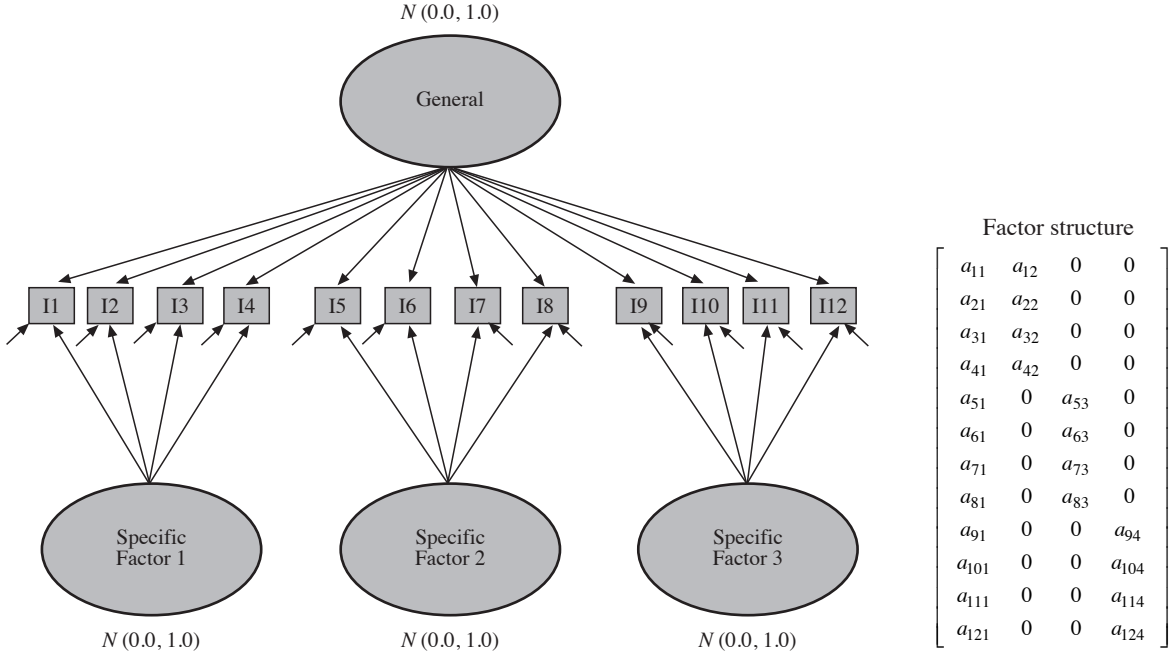

Figure 1. *Bifactor model path diagram*

reasons why the bifactor model has seen little application to non-cognitive measures.

A closely related limitation of the bifactor model is model specification. Reise et al. (2010) advised that for stable parameter estimation one should have at least three group factors, for each group factor there should be at least three indicators, and the number of indicators should be balanced across all group factors. The question then becomes, can I still apply the bifactor model if my theoretical representation is lacking in one of these areas? The answer is "it depends." For one, within a SEM framework one should always have at least three indicators per latent construct for identification purposes. Furthermore, the requirement of possessing at least three specific factors holds true in the second-order model, where it is required that there are at least three first-order factors that load onto the second-order factor (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012). If these first two conditions are not met, one should not apply the bifactor model. In terms of the last condition, having an unequal number of indicators across specific factors will impact reliability estimates of the subscales; however, keeping this mind, one can still fit the model.

Lastly, the bifactor model requires an additional restrictive assumption beyond orthogonality, which is that each indicator load on one general factor *and* one and only one specific factor. Allowing items to cross-load on multiple specific factors would lead to untrustworthy item parameter estimates. Such a restriction on the structure of the multidimensionality may limit the application of the bifactor model. However, this is one of the major reasons why Reise (2012) promoted the use of exploratory bifactor analysis, which allows for indicators to cross-load on specific factors (For a detailed discussion on exploratory bifactor analysis see Jennrich & Bentler, 2011). Such analyses would allow researchers to better understand the structure of the data before applying confirmatory procedures, which is particularly vital with the restrictive assumptions that are inherent in the confirmatory canonical bifactor model.

Model selection. Considering plausible rival hypotheses is an important part of gathering evidence to support the validity of scored-based inferences (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In terms of evidence based on internal structure, rival hypotheses include alternative theoretical models. For example, when a measure is hypothesized to compose one general and multiple specific factors, as is the case with the bifactor model, it is imperative to consider alternative score interpretations. One such alternative hypothesis is that reporting separate scores for the general and specific factors is unnecessary as the score variance can be captured by one prominent dimension. That is, although a model may demonstrate adequate model fit for a multidimensional model, practical and technical considerations (e.g., lack of adequate reliability on the subscales, desire to employ unidimensional IRT applications, etc.) may dictate that reporting a unidimensional model is "good enough" or preferred. In this case, one would be comparing two competing models, the bifactor and unidimensional models. To determine which model best represents the sample data the following four techniques will be discussed: 1) comparison of model fit statistics, 2) ratio of variance accounted for by the general factor over the variance accounted for by the specific factors, 3) the degree to which total scores reflect a common variable, and 4) the viability of reporting subscale scores as indicated by subscale reliability. An empirical example is provided following a discussion of these four techniques.

Traditionally within the SEM framework, model fit statistics are employed to determine the adequacy of a model. For example, to determine the fit of confirmatory models, heuristic guidelines are applied to popular indices, such as CFI, TLI, RMSEA, and SRMR. After obtaining model fit for both unidimensional and bifactor models, one can directly compare the two competing models via the change in CFI (ΔCFI) index as the unidimensional model is hierarchically nested within the bifactor model (Reise, 2012). This index is generally preferred to the traditional Chi-square difference test as ΔCFI has been demonstrated to provide stable performance with various conditions, such as sample size, amount of invariance, number of factors, and number of items (Meade, Johnson, & Braddy, 2008). In contrast, the Chi-square statistic is notoriously known for being highly sensitive to sample size. The ΔCFI is calculated as:

$$\Delta CFI = CFI_{M1} - CFI_{MO} \quad (1)$$

where CFI_{M1} is equal to the CFI value obtained for model 1, and CFI_{MO} is equal to the CFI value obtained for model 0. Based on simulation analyses, Cheung and Rensvold (2002) have recommended that a $\Delta CFI \leq .01$ supports the invariance hypothesis.

This approach for assessing whether data are unidimensional "enough" is quite popular within the SEM framework (Cook & Kallen, 2009). However, such an approach does not shed light on the amount of variance accounted for by the general factor over that accounted for by the specific factors nor does it provide information regarding the viability of reporting a composite score or separate scores on the specific factors. Use of fit indices limits one's assessment of determining the technical adequacy of reporting multidimensional scores that may be adequately represented by a unidimensional model. This assertion is reflected in recent work by Reise, Scheines, Widaman, and Haviland (2013) who have demonstrated that use of fit indices to determine whether data are unidimensional "enough" is not optimal if the data have a multidimensional bifactor structure. This research illustrated that if item response data are bifactor, and those data are forced into a unidimensional model, parameter bias (particularly in structural parameters that depend on loading bias) is a function of the expected common variance (ECV) and percentage of uncontaminated correlations (PUC), whereas model fit indices are a poor indicator of parameter bias. ECV, which provides a ratio of the strength of the general to group factors, is defined as follows:

$$ECV = \frac{\sum_{i=1}^{I_G} \lambda_G^2}{\sum_{i=1}^{I_G} \lambda_G^2 + \sum_{i=1}^{I_{s1}} \lambda_{s_1}^2 + \sum_{i=1}^{I_{s2}} \lambda_{s_2}^2 + \ldots + \sum_{i=1}^{I_{sn}} \lambda_{s_n}^2} \quad (2)$$

where I_G = total number of items loading onto the general factor, I_{s_1} = the number of items loading on specific factor 1, I_{s_2} = the number of items loading on specific factor 2, I_{s_n} = the number of items loading on specific factor n, λ_G^2 = the squared factor loadings of the general factor, $\lambda_{S_1}^2$ = the squared factor loadings of specific factor 1, $\lambda_{S_2}^2$ = the squared factor loadings of specific factor 2, and $\lambda_{S_n}^2$ = the squared factor loadings of specific factor n.

As the ECV value increases to 1, there is evidence to suggest that a strong general dimension is present in the bifactor data. Although this value can be used as an index of unidimensionality,

its interpretation is moderated by PUC. That is, PUC moderates the effects of factor strength on biasing effects when applying a unidimensional model to bifactor data (Reise, Scheines, Widaman, & Haviland, 2013). PUC can be defined as the number of uncontaminated correlations divided by the number of unique correlations:

$$\text{PUC} = \frac{\frac{I_G \times (I_G - 1)}{2} - \left(\frac{I_{s_1} \times (I_{s_1} - 1)}{2} + \frac{I_{s_2} \times (I_{s_2} - 1)}{2} + \ldots + \frac{I_{s_n} \times (I_{s_n} - 1)}{2} \right)}{\frac{I_G \times (I_G - 1)}{2}} \quad (3)$$

where I_G = the number of items loading on the general factor, I_{s_1} =the number of items loading on specific factor 1, I_{s_2} = the number of items loading on specific factor 2, I_{s_n} = the number of items loading on specific factor *n*. When PUC values are very high (>.90), unbiased unidimensional estimates can be obtained even when one obtains a low ECV value (Reise, 2012). More specifically, when the PUC values are very high, the factor loadings of the unidimensional model will be close to those obtained on the general factor in the bifactor model.

In addition to ECV and PUC values, researchers can compute reliability coefficients to determine if composite scores predominately reflect a single common factor even when the data are bifactor. As noted by Reise (2012), the presence of multidimensionality does not dictate the creation of subscales nor does it ruin the interpretability of a unit-weighted composite score. Instead, researchers must make the distinction between the degree of unidimensionality and the degree to which total scores reflect a common variable. This latter assessment can be accomplished by computing coefficient omega hierarchical, which is defined as:

$$\omega_H = \frac{\left(\sum \lambda_{iG}\right)^2}{\left(\sum \lambda_{iG}\right)^2 + \left(\sum \lambda_{iS_1}\right)^2 + \left(\sum \lambda_{iS_2}\right)^2 + \ldots + \left(\sum \lambda_{iS_n}\right)^2 + \sum \theta_i^2} \quad (4)$$

where λ_{iG}= the factor loading for item *i* on the general factor, λ_{iS_1}= the factor loading for item *i* on specific factor 1, λ_{iS_2} = the factor loading for item *i* on specific factor 2, λ_{iS_n} = the factor loading for item *i* on specific factor *n*, and θ_i^2= the error variance for item *i*. Large ω_H values indicate that composite scores primarily reflect a single variable, thus providing evidence that reporting a unidimensional score is viable. Lastly, if this evaluation proves to be inconclusive one can compute the reliability of subscale scores once controlling for the effect of the general factor. This reliability coefficient, which Reise (2012) termed omega subscale (ω_s), can be computed as follows:

$$\varpi_s = \frac{\left(\sum \lambda_{iS_n}\right)^2}{\left(\sum \lambda_{iG}\right)^2 + \left(\sum \lambda_{iS_n}\right)^2 + \sum \theta_i^2} \quad (5)$$

High values indicate that the subscales provide reliable information above and beyond the general factor, whereas low values suggest that the subscales are not precise indicators of the specific factors.

An illustration

To illustrate the basic concepts of using CFA to assess internal structure and model selection, we examined a survey measuring student engagement (SE). The survey was comprised of 27 four-point Likert-type items and was administered to 1,900 participants. Based on theory and previous research, the survey was hypothesized to measure four latent variables: self-management of learning (SML), application of learning strategies (ALS), support of classmates (SC), and self-regulation of arousal (SRA). Nine of the items loaded on SML, ten items loaded on ALS, six items loaded on SC, and three items loaded on SRA. All four latent variables as well as the measurement errors were expected to be uncorrelated in the measurement model. Alternatively, a unidimensional model was also fit to the sample data to determine whether the general student engagement dimension could account for the majority of the score variance. Parameter estimation was conducted in Mplus, version 5 (Muthén & Muthén, 2007) applying the weighted least squares with mean and variance adjustment (WLSMV) estimator to improve parameter estimation with categorical data. Adequate model fit was represented by CFI and TLI values >.95, as well as an RMSEA value <.06 (Hu & Bentler, 1999).

Table 4 provides the standardized factor loading estimates for both the unidimensional and bifactor models. Results demonstrated inadequate model fit to the sample data for the unidimensional model as indicated primarily by a small CFI value, CFI= .80, TLI= .97, and RMSEA= .07. In contrast, model fit was drastically improved when fitting the bifactor model, CFI= .94, TLI= .99, RMSEA= .04, and a ΔCFI index of .14. Examination of the factor loadings (Table 1) demonstrated statistically significant factor loadings of moderate strength for items 7 and 9 on SML, items 10, 12, and 16 on ALS, items 1, 20, 22, and 27 on SC, and all items on SRA. These findings suggest that the specific factors accounted for a significant amount of variance for many of the items above and beyond the variance accounted for by the general factor. Based solely on model fit statistics, one would conclude that the data were not unidimensional "enough" and that a bifactor model best represented the sample data for the models evaluated. However, as mentioned before, model fit statistics do not provide information related to the parameter bias that comes about by representing bifactor data with a unidimensional representation.

The first step in examining parameter bias that is brought about by applying a unidimensional model to bifactor data is to evaluate ECV and PUC. In this example, the sum of the squared factor loadings was 9.13, 0.40, 0.63, 0.91, and 0.73 for the SE, SML, ALS, SC, and SRA factors, respectively (see Table 1). Applying these values to equation 2, ECV was calculated as follows:

$$\text{ECV} = \frac{9.13}{9.13 + 0.40 + 0.63 + 0.91 + 0.73} = .79 \quad (6)$$

The results demonstrated that the ratio of the strength of the general to group factors was .79, which suggested that a very strong general factor was present. However, as mentioned, the interpretation of ECV is mediated by PUC. In this example, the number of unique correlations was [(27×26)/2] = 351. As there were 8, 10, 6, and 3 items that loaded on each specific factor, respectively, the number of correlations for items within group

factors was [((8×7)/2) + ((10×9)/2)+ ((6×5)/2)+ ((3×2)/2)] = 91. Therefore, the number of uncontaminated correlations was 351-91= 260, and the proportion of uncontaminated correlations was 260/351= .74, which is moderate-high with extreme values being represented by anything >.90.

Although the PUC value was not as high as one would hope, its value is dependent on the number of group factors. For example, higher PUC values would be obtained if increasing the number of group factors from 3 to 9, which would produce [((3×2)/2) × 9]= 27 uncontaminated correlations and a proportion of uncontaminated correlations of (324/351)= .92. Nevertheless, in comparing the factor loadings between the general factor from the bifactor model and the unidimensional factor loadings there was a high degree of similarity, r= .88, which demonstrated that the variance accounted for by the general factor was impacted minimally with the inclusion of the specific factors (see Table 1). Such a finding in combination with the ECV and PUC results suggested that a strong general factor was present in the bifactor data.

Table 1
Factor loadings for unidimensional and bifactor models

	Unidimensional	Bifactor					
Item	λ_{SE}	λ_{SE}	λ_{SML}	λ_{ALS}	λ_{SC}	λ_{SRA}	θ^2
1	.59	.52			.53		.45
2	.54	.55	.07				.69
3	.56	.56		.10			.68
4	.53	.50				.71	.25
5	.55	.54		.17			.68
6	.61	.60				.33	.53
7	.61	.59	.47				.43
8	.56	.55		.21			.65
9	.62	.60	.38				.49
10	.57	.54		.36			.58
11	.54	.56		-.03			.69
12	.55	.51		.46			.53
13	.50	.47			.28		.70
14	.55	.53				.34	.61
15	.61	.61		.11			.62
16	.58	.55		.36			.57
17	.60	.59		.18			.62
18	.61	.63	-.04				.61
19	.56	.58	.00				.67
20	.65	.61			.31		.53
21	.65	.65			.07		.57
22	.65	.59			.47		.44
23	.68	.69	.09				.51
24	.65	.66	.06				.56
25	.64	.64	.11				.57
26	.61	.60		.17			.61
27	.68	.62			.48		.39
$(\sum\lambda^2)$		9.13	.40	.63	.91	.73	
$(\sum\lambda)$		244.61	1.30	4.37	4.58	1.90	

Note: λ_{SE} = factor loading for the student engagement factor, λ_{SML} = factor loading for the self-management of learning factor, λ_{ALS} = factor loading for the application of learning strategies factor, λ_{SC} = factor loading for the support of classmates factor, λ_{SRA} = factor loading for the self-regulation of arousal factor, and θ^2= item residual variance (only reported for bifactor model due to reliability coefficient calculations)
<.05

The next step was to evaluate the degree to which a total score reflected a common variable. This was accomplished by first computing the squared sums of the factor loadings, which were 244.61, 1.30, 4.37, 4.58, and 1.90 for the SE, SML, ALS, SC, and SRA factors, respectively. In addition, the sum of the residual variance across all 27 items was equal to 15.23. These values were then applied to equation 5 as follows:

$$\varpi_H = \frac{244.61}{244.61+1.30+4.37+4.58+1.90+15.23} = .90 \quad (7)$$

The results demonstrated an omega hierarchical of .90, which suggested that a very high amount of the variance in summed scores could be attributed to the single general factor. The last step of the analysis was to compute the reliability of the subscales by controlling for the general factor variance. The omega subscale reliabilities were calculated for the four specific factors as follows:

$$\varpi_{\text{SML}} = \frac{1.30}{244.61+1.30+15.23} = .004 \quad (8)$$

$$\varpi_{\text{ALS}} = \frac{4.37}{244.61+4.37+15.23} = .02 \quad (9)$$

$$\varpi_{\text{SC}} = \frac{4.58}{244.61+4.58+15.23} = .02 \quad (10)$$

$$\varpi_{\text{SRA}} = \frac{1.90}{244.61+1.90+15.23} = .007 \quad (11)$$

As can be seen, the reliabilities of the scores for the specific factors after controlling for the variance accounted for by the general factor were extremely low. Such low reliability estimates demonstrate that reporting scores on the specific factors would provide unreliable information.

In summarizing the results of assessing the unidimensional and bifactor models tested in this example, one would conclude that although unique factors associated with the individual scales were present, the information that they provided was of negligible consequence. That is, from a practical standpoint, reporting multidimensional scores would be invalid as the technical adequacy was lacking, due to a strong general factor, a high amount of variance being accounted for in summed scores by the general factor, and extremely low reliability estimates for scores on the specific factors. This example demonstrates the need for researchers to go beyond the use of model fit statistics in deciding whether to employ a multidimensional representation as often a unidimensional model can be more adequate.

Assessing measurement invariance

One societal concern related to measurement is the lack of test fairness for distinct subgroups within the population. Although the evaluation of fairness incorporates legal, ethical, political, philosophical, and economic reasoning (Camilli, 2006), from

a psychometric perspective, one can define fairness as a lack of systematic bias (measurement invariance). Bias is a technical term that comes about when there are systematic deficiencies in the test that lead to differential interpretation of scores by subgroup. From this perspective, the main concern in evaluating bias is to determine whether knowledge of an examinee's group membership influences the examinee's score on the measured variable (e.g., an item, subdomain, or test), given the examinee's status on the latent variable of interest (Millsap, 2011). If group membership is found to impact score-based inferences, one would conclude that the measure contains construct-irrelevant variance. If not, one would conclude that the measure demonstrates equivalence (invariance) across subgroups. Therefore, for a test to be fair (from a psychometric perspective) one must demonstrate measurement invariance across all distinctive subgroups being evaluated. This assertion is reflected in Standard 7.1 of the 1999 *Standards*, which states:

> "…the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 80).

There are numerous statistical approaches for assessing measurement invariance. These methods can be categorized into three distinctive groups: 1) linear measurement models, 2) non-linear measurement models, and 3) observed score methods (Millsap, 2011). Furthermore, these approaches can be broken down into methods that examine invariance at the scale- and item-levels (Zumbo, 2003). Scale-level analyses are primarily concerned with the degree of invariance observed within common factor analytic models across groups. In contrast, item-level analyses (differential item functioning (DIF)) evaluate invariance for each item individually. The literature on DIF is extensive and spans more than 40 years. As a result, the main focus of this section will be on describing the multiple-group confirmatory factor analytic method for assessing invariance at the scale-level. For a general introduction to DIF, as well as the various methods available for analyzing item-level invariance, the reader is referred to Sireci and Rios (2013).

Multiple Group Confirmatory Factor Analysis (MGCFA)

MGCFA is a theory-driven method used to evaluate formal hypotheses of parameter invariance across groups (Dimitrov, 2010). MGCFA is advantageous to use when establishing construct comparability as it allows for: 1) simultaneous model fitting across multiple groups, 2) various levels of measurement invariance can be assessed, 3) the means and covariances of the latent constructs' are disattenuated (i.e., controls for measurement error), and 4) direct statistical tests are available to evaluate cross-group differences of the estimated parameters (Little & Slegers, 2005). Conducting MGCFA requires a number of hierarchical steps, which depend on the desired inferences that the researcher is interested in. These hierarchical steps can be described as first establishing a baseline model separately for each group, and then systematically evaluating hierarchically nested models to determine the level of invariance present across groups. This systematic process is known as *sequential constraint imposition* as model parameters across groups are allowed to be freely estimated with greater constraints on the parameters being placed as adequate model fit for less restricted models is obtained. Comparison of hierarchically nested models can be conducted via the ΔCFI index.

Levels of measurement invariance

There are various levels of invariance; however, for the purposes of this paper, we will only discuss configural, metric, scalar, and strict factorial invariance; however, it should be noted that there are other forms of equivalence, such as invariance of item-uniqueness (See Dimitrov, 2010). The most basic and necessary condition for group comparisons is configural invariance, which assesses whether there is conceptual equivalence of the underlying variable(s) across groups (Vandenberg & Lance, 2000). From a factor analytic perspective, configural invariance is reflected in the use of identical indicators to measure the same latent construct(s) of interest across groups. A more restrictive form of invariance is metric equivalence, which assumes both configural invariance and equivalent strengths between the indicators and latent variable (factor loadings) across groups. Attainment of metric equivalence denotes equal measurement units of the scale designed to measure the latent construct across groups. This form of equivalence allows for indirect comparisons as the score intervals are equal across groups but the measurement units do not share the same origin of the scale. As a result, direct comparisons of group means are not valid. To make direct comparisons of latent group means, it is necessary to attain scalar equivalence. This form of invariance subsumes both configural and metric equivalence, as well as assumes that the scales of the latent construct possess the same origin, which is indicated by equal intercepts across groups. Lastly, when one is concerned with the equivalence of covariances among groups for a number of latent factors within the model, strict factorial invariance is of interest. For a detailed example of conducting a scale-level measurement invariance analysis, the reader is referred to Dimitrov (2010).

Reliability: Internal consistency

Internal consistency reliability represents the reproducibility of test scores on repeated test administrations taking under the same conditions and is operationally defined as the proportion of true score variance to total observed score variance (Crocker & Algina, 1986). Although there are several methods for estimating reliability of a composite or subscale score such as split-half reliability, coefficient α (Cronbach, 1951) is arguably the most commonly used statistic. Cronbach (1951) demonstrated that coefficient α is the average of all possible split-half reliability values for a test and is computed as follows:

$$\hat{\alpha} = \frac{I}{I-1}\left(1 - \frac{\sum_{i=1}^{I} s_i^2}{s_x^2}\right) \tag{12}$$

I represents the number of items; s_i^2 represents the variance of scores for item i; and s_x^2 represents the test score variance.

Despite the widespread use of coefficient α, it is not without its limitations. For example, in most cases when the measurement errors are uncorrelated (except for the tau-equivalent condition), coefficient α will often underestimate reliability (Crocker & Algina, 1987; Lord & Novick, 1968). When the measurement errors are correlated, for example due to method effects or items that share a common stimulus, coefficient α can either underestimate or overestimate reliability (Raykov, 2001). To address these limitations, CFA can be used to provide a more accurate estimate of reliability. Reliability can be estimated from the parameter estimates in a CFA model as follows:

$$\rho_Y = \frac{\left(\sum_i \lambda_i\right)^2}{\left(\sum_i \lambda_i\right)^2 + \sum_i VAR(\delta_i) + 2\sum_{i,j} COV\left(\delta_i, \delta_j\right)} \quad (13)$$

λ represents the unstandardized factor loading; $VAR(\delta_i)$ represents the measurement error variance; and $COV(\delta_i,\delta_j)$ represents the covariance in measurement errors. Essentially, the numerator represents true score variance and equals the squared sum of the unstandardized factor loadings. The denominator represents the total observed score variance and includes the true score variance, error variance and any non-zero correlated measurement errors.

To illustrate how to compute reliability using a CFA model, we utilized the factor loadings for a six item subscale (see Table 2). The reliability for a subscale can be computed using the model parameter estimates. For example, for this one subscale, the true variance equals the squared sum of the unstandardized factor loadings:

$$(1.01 + 0.80 + 1.02 + 0.91 + 1.06 + 1.12)^2 = 24.11 \quad (14)$$

The total variance of the subscale is

$$24.11 + 0.53 + 0.70 + 0.51 + 0.61 + 0.47 + 0.41 = 26.81 \quad (15)$$

Therefore, the reliability estimate based on the CFA model is 24.11/26.81 = 0.90. In comparison to coefficient α, which equaled 0.80 for the subscale, the reliability estimate based on the CFA model parameter estimates was larger most likely because the tau equivalence condition was not met.

Table 2
Factor loadings to illustrate reliability estimation

Item	Unstandardized λ	Standardized λ
1	1.01	0.70
2	0.80	0.53
3	1.02	0.71
4	0.91	0.62
5	1.06	0.73
6	1.12	0.76

Conclusion

The need to gather evidence that supports the validity of score-based inferences is imperative from scientific, ethical, and legal perspectives. In this article we provided a general review of methodological procedures to evaluate one form of validity evidence, internal structure, by specifically focusing on assessment of dimensionality, measurement invariance, and reliability within a factor analytic framework. In addition, an overview of the bifactor model, as well as techniques that go beyond fit indices for determining model selection, was illustrated. The methods outlined in this paper, when applied appropriately, will assist researchers in gathering evidence to strengthen the validity of intended scored-based inferences.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *The standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

Arifin, W.N., Yusoff, M.S.B., & Naing, N.N. (2012). Confirmatory factor analysis (CFA) of USM Emotional Quotient Inventory (USMEQ-i) among medical degree program applicants in Universiti Sains Malaysia (USM). *Education in Medicine Journal, 4*(2), e1-e22.

Bollen, K.A. (1989). *Structural equation models with latent variables*. New York, NY: Wiley.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Camilli, G. (2006). Test Fairness. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). Westport, CT: American Council on Education/Praeger.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Chen, F.F., Hayes, A., Carver, C.S., Laurenceau, J.P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80*(1), 219-251.

Chen, F.F., Sousa, K.H., & West, S.G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*(3), 471-492.

Chen, F.F., West, S.G., & Sousa, K.H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Cook, K.F., & Kallen, M.A. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumptions. *Quality of Life Research, 18,* 447-460.

Dimitrov, D.M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149.

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement, 35*(8), 604-622.

Holzinger, K.J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41-54.

Assessing stress in cancer patients: A second-order factor analysis model for the Perceived Stress Scale. *Assessment, 11*(3), 216-223.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Jennrich, R.I., & Bentler, P.M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*(4), 537-549.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38*(1), 32-60.

Kenny, D.A. (1976). An empirical application of confirmatory factor analysis to the multitraite-multimethod matrix. *Journal of Experimental Social Psychology, 12*, 247-252.

Kline, R.B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Krause, N., & Hayward, R.D. (2013). Assessing stability and change in a second-order confirmatory factor model of meaning in life. *Journal of Happiness Studies*, 1-17.

Levant, R.F., Hall, R.J., & Rankin, T.J. (2013). Male Role Norms Inventory-Short Form (MRNI-SF): Development, confirmatory factor analytic investigation of structure, and measurement invariance across gender. *Journal of Counseling Psychology, 60*(2), 228-238.

Li, Y., & Lissitz, R.W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36*(1), 3-20.

Little, T.D., & Slegers, D.W. (2005). Factor analysis: Multiple groups with means. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 617-623). Chichester, UK: Wiley.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement, 36*(8), 670-688.

Marsh, H.W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335-361.

Martel, M.M., Roberts, B., Gremillion, M., von Eye, A., & Nigg, J.T. (2011). External validation of bifactor model of ADHD: Explaining heterogeneity in psychiatric comorbidity, cognitive control, and personality trait profiles within DSM-IV ADHD. *Journal of Abnormal Child Psychology, 39*(8), 1111-1123.

Meade, A.W., Johnson, E.C., & Braddy, P.W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, L.K., & Muthén, B.O. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Reise, S.P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667-696.

Reise, S.P., Moore, T.M., & Haviland, M.G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544-559.

Reise, S.P., Scheines, R., Widaman, K.F., & Haviland, M.G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5-26.

Salgueiro, M.F., Smith, P.W.F., & Vieira, M.D.T. (2013). A multi-process second-order latent growth curve model for subjective well-being. *Quality & Quantity: International Journal of Methodology, 47*(2), 735-752.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187.

Thomas, M.L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological Assessment, 24*(1), 101-113.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association.

Vandenberg, R.J., & Lance, C.E. (2000). A review of synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70.

Wang, M., Schalock, R.L., Verdugo, M.A., & Jenaro, C. (2010). Examining the factor structure and hierarchical nature of the quality of life construct. *American Journal on Intellectual and Developmental Disabilities, 115*(3), 218-233.

Wolf, E.J., Miller, M.W., & Brown, T.A. (2011). The structure of personality disorders in individuals with posttraumatic stress disorder. *Personality Disorders: Theory, Research, and Treatment, 2*(4), 261-278.

Xie, J., Bi, Q., Li, W., Shang, W., Yan, M., Yang, Y., Miao, D., & Zhang, H. (2012). Positive and negative relationship between anxiety and depression of patients in pain: A bifactor model analysis. *Plos ONE, 7*(10).

Zumbo, B.D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*(2), 136-147.