Conducting Measurement Invariance Tests with Ordinal Data: A Guide for Social Work Researchers

Natasha K. Bowen *University of North Carolina at Chapel Hill*
Rainier D. Masa *University of North Carolina at Chapel Hill*

ABSTRACT ***Objective:*** The validity of measures across groups is a major concern for social work researchers and practitioners. Many social workers use scales, or sets of questionnaire items, with ordinal response options. However, a review of social work literature indicates the appropriate treatment of ordinal data in measurement invariance tests is rare; only 3 of 57 articles published in 26 social work journals over the past 12 years used proper testing procedures. This article synthesizes information from the literature and provides recommendations for appropriate measurement invariance procedures with ordinal data. ***Method:*** We use data from the Cebu Longitudinal Health and Nutrition Survey to demonstrate applications of invariance testing with ordinal data. Using a robust weighted least squares estimator and polychoric correlation matrix, we examine invariance of a 10-item Perceived Stress Scale (PSS) across 2 young adult groups defined by health status. We describe 2 competing approaches: a 4-step approach, in which factor loadings and thresholds are tested and constrained separately; and a 3-step approach, in which loadings and thresholds are tested and constrained in tandem. ***Results:*** Both approaches lead to the same conclusion that the 2 dimensions of the PSS are noninvariant across health status. In the absence of invariance, mean scores on the PSS factors cannot be validly compared across groups, nor should latent variables be used in the hypothesis testing across the 2 groups. Readers are directed to online resources. ***Conclusions:*** Careful examination of social work scales is likely to reveal fit or noninvariance problems across some groups. Use of appropriate methods for invariance testing will reduce misuse of measures in practice and improve the rigor and quality of social work publications.

KEYWORDS: Measurement invariance, Scale development, Statistical factor analysis, Ordinal data, Stress scale, Psychometrics

doi: 10.1086/681607

Because social work researchers and practitioners work with populations that vary by age, race/ethnicity, gender, educational attainment, disability, cognitive functioning, physical and mental health status, and other characteris-

Journal of the Society for Social Work and Research, Volume 6, Number 2. 2334-2315/2015/0602-0003/$10.00.

tics, the validity of measures across groups is a major concern. Social workers commonly use scales, or sets of questionnaire items, to measure complex attitudes, perceptions, behaviors, and other constructs. Scales assume that levels of an underlying phenomenon or construct cause respondents to choose certain responses to scale items. When scores collected from different populations are interpreted the same way (e.g., with the same eligibility cutoff for services), social workers are assuming that identical scores represent the same level of the construct for members of different groups. However, the nature and magnitude of relationships between items and a latent phenomenon may differ across groups, meaning their scores should not be interpreted the same (Dimitrov, 2010; Muthén & Asparouhov, 2002; Raykov, Marcoulides, & Millsap, 2013; Sass, 2011). Being aware of differences in scale performance across groups is critical for both practice and research. Without knowledge of differences, clinicians might deny services to members of a subgroup because their assessment scores are below a clinical cutoff despite high levels of impairment. Researchers might draw erroneous conclusions about relationships among social, emotional, or behavioral constructs and outcomes for subgroups. Their conclusions could translate into guidelines for intervention that are inappropriate for some clients.

Tests of cross-group similarities in the relationships between latent constructs and scale items are tests of measurement invariance. Multiple approaches exist for testing measurement invariance, but we focus on confirmatory factor analysis (CFA), which is a type of structural equation modeling (SEM). This article assumes readers have a basic knowledge of SEM and CFA and the logic of hierarchical model comparisons to identify the best model. Tests of measurement invariance are simply multiple group SEM analyses that focus on the measurement parameters of a model instead of predictive relationships among latent variables. In multiple group tests of substantive parameters, researchers are often looking for statistically significantly different parameter estimates based on hypothesized group differences. In contrast, when they conduct tests of measurement invariance, researchers most often hope to find non significant differences. The implications of noninvariance in scales for practice and research are discussed later.

Social work researchers conducting CFA often use maximum likelihood (ML), the default estimator in most SEM programs, and analyze a covariance matrix of their scale items. Because these analysis options are not appropriate for the most common type of social work data—data collected with ordinal questionnaire items—we focus on the procedures currently considered most appropriate for invariance testing with ordinal variables. The goal of this article is to equip social work researchers with knowledge and skills to conduct and publish high-quality measurement invariance studies of ordinal items using CFA. We refer to Mplus (Muthén & Muthén, 1998–2014), one SEM program that accommodates common problems of social work data, including ordinality.

Background and Significance

General Invariance Testing

Invariance testing involves comparing the fit of a succession of nested models, each with more equality constraints on parameters across groups than the previous model (Dimitrov, 2010). The default ML and covariance matrix options of most SEM programs are appropriate for normally distributed, continuous variables. Measurement parameters of interest under these conditions are factor loadings, intercepts, and residual variances (Vandenberg & Lance, 2000).

Although numerous approaches to invariance testing have been described in the literature (Vandenberg & Lance, 2000), certain steps are common across descriptions. As a first step, many scholars recommend identifying a baseline model for each group by conducting a CFA for one group at a time. Ideally, the same baseline model is confirmed for all groups; however, minor differences can be modeled (Byrne, Shavelson, & Muthén, 1989). In practice, it also appears that baseline models do not have to fully meet prespecified fit criteria (Byrne et al., 1989; Raykov, Marcoulides, & Li, 2012). After finding the best model for each group, the actual invariance testing begins.

The first level of invariance *is configural* invariance. A factor model with configural or *form* invariance has the same pattern of factor loadings across groups (Dimitrov, 2010); that is, the same items load onto the same factors across groups. No cross-group constraints are placed on model parameters beyond fixing the first loading of each factor to 1.0 for identification purposes. Fixing a loading at 1.0 sets the variance of the latent variable equivalent to the variance of the common or shared variance of the associated indicator (Steiger, 2002). The indicator whose loading is fixed to 1.0 is called the *referent indicator* (Johnson, Meade, & DuVernet, 2009) or simply the *referent* (French & Finch, 2006). The goal of the configural test is to determine if the unconstrained multiple group model meets fit criteria (Millsap & Olivera-Aguilar, 2012). If it does, the configural model becomes the model with which subsequent models are compared. If the unconstrained configural model does not meet minimal fit criteria, then invariance testing does not proceed because the hypothesized factor model is not acceptable for one or both groups.

The second level of invariance is *metric* or *weak* invariance. Scales with metric invariance have statistically equivalent factor loadings across groups (Dimitrov, 2010) in addition to configural invariance. Non referent loadings are constrained to be equal and model fit is compared to the fit of the configural model. Noninvariant loadings signify that indicators have different relationships (e.g., are more or less important) to the latent variable across groups, or the latent construct is defined differently across groups (Muthén & Asparouhov, 2002; Sass, 2011). Similar to configural invariance, metric invariance is not sufficient to justify equivalent interpretation of scale scores across groups in practice or research.

The third level of invariance is *scalar* or *strong* invariance. For scales with continuous indicators, scalar invariance is defined by the presence of invariant intercepts (i.e., in the equations relating latent variables to observed item scores) in addition to invariant loadings and the same pattern of item loadings on factors. Scales with this level of invariance are considered adequately invariant for most practice and research purposes. Scalar invariance implies that differences in scale scores are caused by differences in true levels of the underlying construct, not other causes (Millsap & Olivera-Aguilar, 2012). If scalar invariance is demonstrated, researchers can compare factor means, variances, and covariances across groups, and can test hypothetical directional relationships among factors in theoretical models (Dimitrov, 2010).

A fourth level of invariance is called *strict or uniqueness invariance*, in which residual variances are equivalent across groups in addition to factor structure, loadings, and intercepts. Strict invariance is not considered necessary for most social work practice and research purposes, so researchers do not usually proceed to this step.

Models with equality constraints almost invariably have worse fit than models in which corresponding parameters are freely estimated for each group. However, the logic behind invariance testing holds that if the decrement in fit is not statistically significant, then the parameter estimates can be considered invariant and constrained to be equal. Often, the change in $\chi^2(\Delta\chi^2)$ per degree of freedom (df) is used to evaluate whether fit has deteriorated significantly, but change in other fit indices have also been recommended (Cheung & Rensvold, 2002; French & Finch, 2006). Scholars continue to study the performance of various fit indices for invariance testing under different conditions (Chen, 2007; Sass, Schmitt, & Marsh, 2014).

Implications of Noninvariance

Invariance tests often do not culminate with an all or nothing verdict about measurement invariance. Byrne et al. (1989) described the logic and methods of testing for partial invariance. Researchers who find a statistically significant change in fit when all the parameters tested in a particular step (e.g., all factor loadings) are constrained to be equal across groups proceed systematically to test individual parameters or subsets of parameters to identify the source of the noninvariance (Millsap & Yun-Tein, 2004). Many researchers consider partial measurement invariance acceptable if the proportion of noninvariant parameters to all parameters tested is small (Dimitrov, 2010; Millsap & Olivera-Alguilar, 2012; Muthén & Asparouhov, 2002; Sass, 2011). However, no definitive definition of "small" is available. Dimitrov suggested: "less than 20% freed parameters seems acceptable in practical applications" (p. 127), but he also stated that researchers should ultimately choose

their own cutoffs. Scholars who suggest a small number of noninvariant parameters is acceptable in a scale believe that when the majority of measurement parameters are equivalent across groups, scale scores can be treated as if they are invariant (Millsap & Olivera-Alguilar, 2012). Researchers using the scales in general SEMs, would allow noninvariant parameters to vary across groups in their models to obtain the most valid score for each group. However, because differences in the definition of the latent variable were minor, latent variable scores could be interpreted equivalently across groups. The literature seems to suggest that with relatively small levels of noninvariance, researchers or practitioners who compute observed composite scores could also ignore the noninvariance (Millsap & Olivera-Alguilar, 2012).

One response to noninvariance used by researchers seeking to create invariant scales is to delete items with noninvariant parameters (Godfrey et al., 2012). However, some invariance scholars (Cheung & Rensvold, 1998, 1999) contend that deleting items might remove the most important information about group differences. When true differences in constructs exist across groups, removing affected items has the potential to make the scale less valid as a measure for one or more of the groups tested. Deletions might also reduce the adequacy of domain sampling of the scale in general. Therefore, we do not recommend item deletion unless these issues have been considered.

For scales with large amounts of noninvariance, the implications are more clear: factor scores and composite scores cannot be interpreted as if the scores have the same meaning across groups. The meaning of the construct for each group should be described and separate statistical models should be run. Multiple group tests of substantive models cannot be performed because factor scores for different groups do not convey equivalent information about underlying factors, or such scores may represent qualitatively different underlying factors.

Throughout invariance testing, researchers must also be attentive to the overall fit of the model. The deterioration of fit statistics to values below the researchers' prespecified fit criteria signifies the specified model does not fit the data well, and the scale is inadequate for use in research or practice.

Invariance Testing with Ordinal Data

Invariance tests for ordinal data are different from those used with continuous variables in terms of the estimator used, the analysis matrix, and the parameters examined. Ordinal variables have response options that have a logical order, such as the range from *strongly disagree* to *strongly agree*; or from *never*, *sometimes*, *often,* to *always*. These options can be logically ordered and by convention are assigned consecutive whole number values. However, because ordinal responses do not correspond to true quantitative values (such as, *0 times per week, 5 times per week,* and

10 times per week), the assignment of numbers to ordinal responses for use in analyses is partially arbitrary. For example, the same five ordered response options could be assigned values 0 to 4, 1 to 5, or 5 to 1 in a dataset or for a particular analysis. Therefore, SEM experts agree that ordinal data should not be analyzed as if the data were continuous (Bollen, 1989; Jöreskog, 2005). "Means, variances, and covariances of ordinal variables have no meaning" (Jöreskog, p. 1). However, as described below, the ordered ranking of responses can be handled using special procedures.

The currently recommended estimator for ordinal data is weighted least squares (WLS; Beauducel & Herzberg, 2006; Bollen, 1989; Jöreskog, 2005; Muthén & Muthén, 1998–2012). WLS estimators make fewer assumptions than ML about variable distributions (Bollen, 1989), which is important because ordinal variables often have non normal distributions. Moreover, studies (e.g., Beauducel & Herzberg, 2006; Flora & Curran, 2004) have suggested that *robust* WLS estimators are the best choice for analyses with ordinal data because they use a diagonal weight matrix, which reduces sample size requirements and prevents certain convergence problems (Bovaird & Koziol, 2012). In Mplus, the recommended robust WLS estimator is a means and variance adjusted WLS labeled as WLSMV (Flora & Curran, 2004; Muthén & Muthén, 1998–2012).

In addition to WLS estimation, a polychoric correlation matrix is used in the analysis of ordinal data instead of the usual covariance matrix (Bollen, 1989; Jöreskog, 2005; Muthén & Muthén, 1998–2012). A polychoric correlation is computed for each pair of ordinal variables in the analysis based on a theoretical assumption that a normally distributed, latent continuous variable underlies the observed frequencies of the observed ordinal responses of each variable (Bollen, 1989; Jöreskog, 2005). According to this theory, each observed ordinal response value corresponds to a range of normalized values between thresholds, or cutoffs, on the underlying latent continuous variable. Use of a polychoric correlation matrix addresses the ordinality of observed variables. Thresholds (or taus; τ_1 to τ_4 in Figure 1a) are cutoffs that divide the underlying normal distributional into five sections, each of which corresponds to an observed ordinal score.

When indicators are ordinal, the parameters of interest in invariance testing are factor loadings (λs, lambdas), thresholds (τs, taus), and residual variances (Vandenberg & Lance, 2000). The levels of invariance are the same as the levels described in the earlier section on general invariance testing, but thresholds are the focus of scalar invariance tests instead of intercepts. Figure 1b, in conjunction with Figure 1a, illustrates the meaning of noninvariant thresholds. Some of the threshold values are different across the two groups. Invariant thresholds would be statistically the same for both groups. Scalar invariance, loading and threshold invariance, is still the minimum required for being able to interpret scores equivalently across groups.

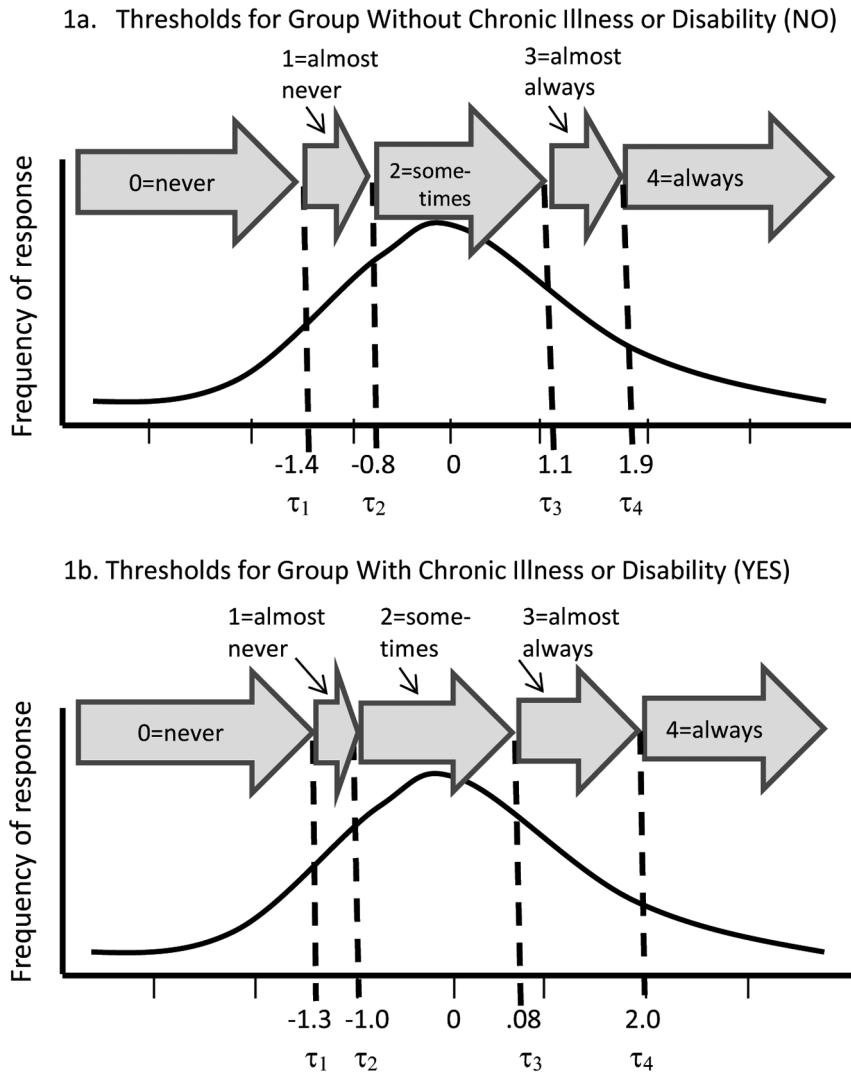

Figure 1. Illustration of noninvariant thresholds across groups with and without chronic illness or disability for the indicator, "In the last 4 weeks, how often have you been upset because of something that happened unexpectedly?" In a threshold model, ranges of normalized scores from underlying continuous latent variables correspond to ordinal response options. The ranges are defined by thresholds (τ or taus, long vertical dashed lines in the illustration). Five response values (0, 1, 2, 3, 4) have four thresholds. Thresholds illustrated in Figure 1b are significantly different from those in 1a. For example, a larger range of true scores correspond to *almost always* in 1b (.08 to 2.0 versus 1.1 to 1.9 in 1a). For NO group members in 1a, a true score of 1 on the latent factor leads to response of *sometimes* on the indicator. For YES group members in 1b, the same true score leads to the response *almost always*.

Ordinal variables typically have multiple thresholds (the number of response options minus one); therefore, ordinal invariance tests usually include more parameters than analogous tests with continuous variables. The presence of thresholds also leads to new issues of identification in CFA models (Muthén & Asparouhov, 2002). Certain thresholds must be constrained to be equal even when a free-threshold model is estimated (Muthén & Muthén, 1998–2012). Specifically, in addition to the referents fixed to 1, the bolded τs in Figure 2 must be constrained to equality across groups.

Unresolved Issues in Invariance Testing with Ordinal Data

A recent analysis identified limitations of all commonly used fit measures (Sass et al., 2014), but concluded that $\Delta\chi^2$ may be the best choice with robust WLS estimation (Sass, 2011; Sass et al., 2014). Other scholars have noted the performance of fit indices for invariance tests with categorical or ordinal data has not been adequately studied (Bovaird & Koziol, 2012).

In addition, the literature indicates use of varying practices and chronicles the ongoing discussion about whether each indicator's loading and thresholds should be constrained and freed simultaneously (Lubke & Muthén, 2004; McLarnon & Carswell, 2013; Millsap & Yun-Tein, 2004; Muthén & Asparouhov, 2002; Sass, 2011; Webber, 2014; Wegmann, 2014). If loadings and thresholds are tested as a set, there is no separate test for invariant factor loadings. Tests proceed from an examination of configural invariance to the examination of scalar invariance. As Sass (2011) pointed out, a researcher could argue loadings and thresholds should be constrained and freed together because they jointly define item functioning; or a researcher could argue that because loadings and thresholds contribute different information about item functioning, they can be constrained and freed separately. Constraining and freeing loadings and thresholds in tandem is suggested in the latest *Mplus User's Guide* (Muthén & Muthén, 1998–2012). However, testing loadings and thresholds separately offers a number of advantages. For example, being able to identify individual noninvariant thresholds and loadings enables researchers to pinpoint and interpret sources of noninvariance (Lubke & Muthén, 2004; Muthén & Asparouhov, 2002; Webber, 2014; Wegmann, 2014). In addition, freeing only individual noninvariant thresholds and loadings has the advantage of reducing the number of parameters modeled as noninvariant, making it easier to satisfy Dimitrov's (2010) "fewer than 20%" guideline.

Another unresolved issue is how to choose the referent indicator. The choice can matter because if a noninvariant loading is chosen, constraining the noninvariant loading to the same value (1.0) across groups will not only hurt overall model fit, but could also affect the results of invariance tests of other parameters (Johnson et al., 2009). Currently, the discussion of choosing referents focuses on analyses of continuous data, so we cannot make recommendations. In the example

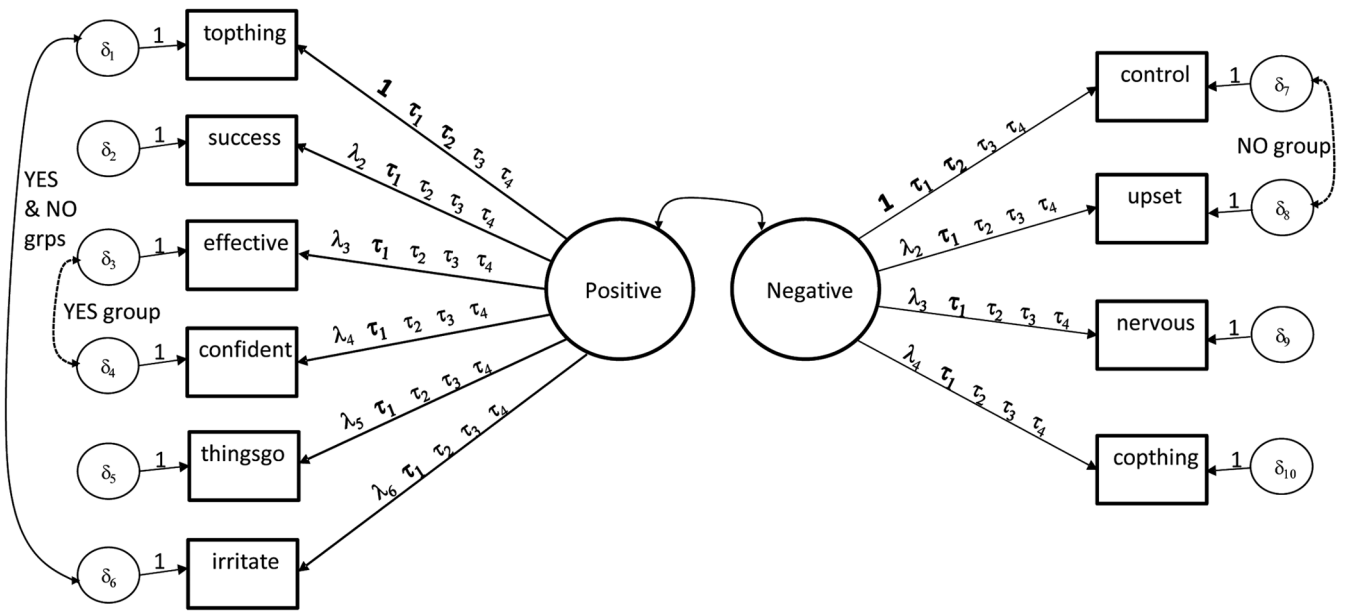

Figure 2. Models tested in invariance testing example: Two-factor Perceived Stress Scale (PSS) model with group-specific baseline correlated errors Note. **λ** = factor loading; τ = threshold; Bolded **1**s indicate fixed factor loadings. Bolded **τ**s indicate thresholds that must be constrained to be equal across groups for identification purposes in metric and scalar models. These fixed and constrained parameters are not tested for invariance. The baseline model for each group had one common correlated error (between δ_1 and δ_6) and one unique correlated error (between δ_3 and δ_4 for the YES group, and between δ_7 and δ_8 for the NO group).

below, we allow Mplus to use its default procedure of making the first indicator listed for a factor the referent. Please look for updates on this topic under the Structural Equation Modeling section at http://ssw.unc.edu/sswsig/ResearchMethods

Invariance Testing of Ordinal Data in the Social Work Literature

To ascertain the need for a guide on proper measurement invariance procedures with ordinal data, we searched issues of 26 primarily quantitative social work journals examined in previous systematic reviews (Sellers, Mathieson, Smith, & Perry, 2006; Thyer, 2008) and published between January 2002 and May 2014. We searched for articles reporting on CFA invariance studies. Our search resulted in a sample of 57 articles, which appeared in 14 of the 26 journals (53.8%). More than half (54.4%) of the invariance reports appeared in either *Research on Social Work Practice* (n = 19; 33.3%) or *Social Work Research* (n =12; 21.1%). Of the 57 articles, only three (Godfrey et al., 2012; Granillo, 2012; Silver Wolf, Dulmus, Maguin, & Fava, 2014) provided adequate information on analytic procedures, reported using robust WLS and a polychoric correlation matrix, and used recommended invariance testing steps. Our review suggests the need among social work researchers for guidance in appropriate invariance testing with ordinal data. More detail about the literature review is available from the second author.

Recommended Steps for Invariance Testing with Ordinal Data

Figure 3 summarizes our literature-based recommendations for measurement invariance testing with ordinal data. The steps in Figure 3 take into account the fact that the findings can be affected by the order in which parameters are tested (Byrne et al., 1989). Specifically, we recommend comparing individual constraints or set of constraints within Steps 3 and 4 to the same less restrictive model. For example, even when an individual factor loading is found to be invariant, that loading should temporarily be freed while the search for other invariant loadings continues. However, researchers should constrain all confirmed invariant loadings before moving on to testing thresholds. When multiple constraints are imposed simultaneously, a significant increase in χ^2 can be caused by a single noninvariant parameter in the set tested or by more than one parameter. Therefore, all loadings and thresholds that are not fixed for identification purposes need to be tested in turn.

An Example of Invariance Testing with Ordinal Data

Source of Data and Measures

In this section, we demonstrate the application of invariance testing with ordinal data from the 2005 follow-up survey of the Cebu Longitudinal Health and Nutrition Survey (CLHNS). CLHNS includes items related to health, demographic,

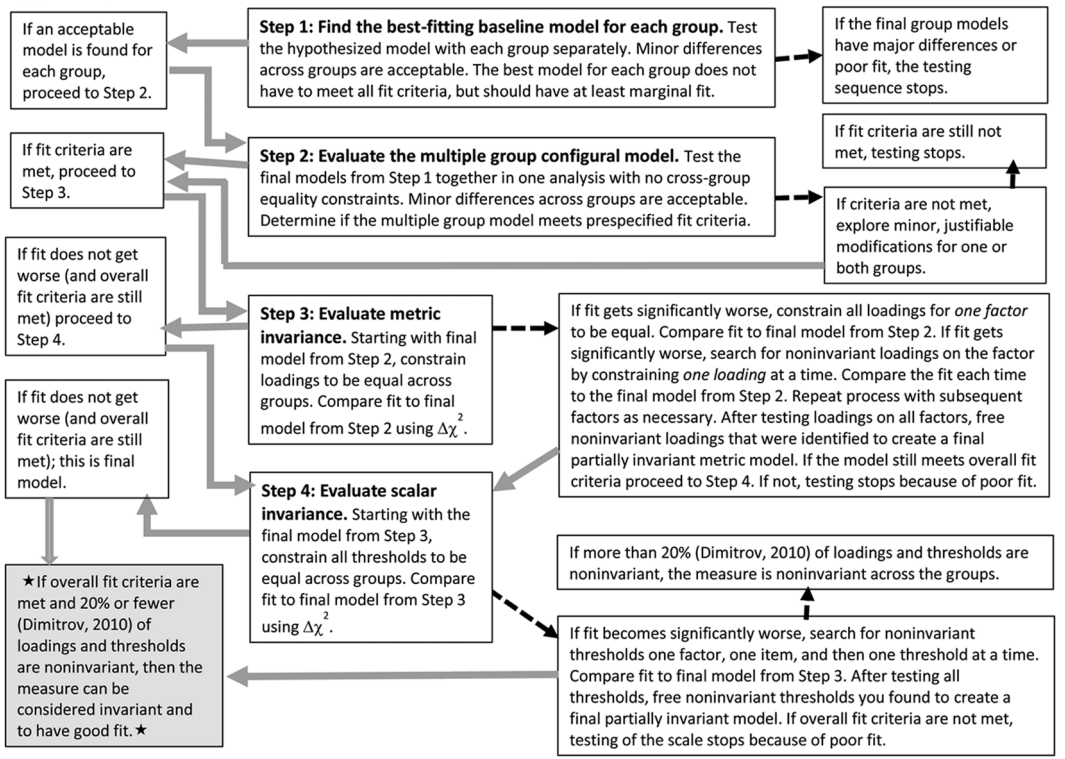

Figure 3. Recommended measurement invariance steps using robust WLS estimation and a polychoric correlation matrix for ordinal data

nutritional and socioeconomic outcomes of Filipina mothers and their children. In 2005, a sample of 1,912 young adults, aged 20 to 22 years, completed numerous health measures, including the 10-item Perceived Stress Scale (PSS; Cohen, Kamarck, & Mermelstein, 1983). Items on the PSS have five response options: *never* (= 0), *almost never* (= 1), *sometimes* (= 2), *fairly often* (= 3) and *very often* (= 4). Sample items include, "In the last 4 weeks, how often have you been upset because of something that happened unexpectedly?" and "In the last 4 weeks, how often have you felt that you were on top of things?" The dataset used for this example is publicly available at http://www.cpc.unc.edu/projects/cebu. Annotated syntax for the following 4-step and 3-step examples is available at Structural Equation Modeling at http://ssw.unc.edu/sswsig/ResearchMethods. Also available at the website is information on Mplus' "shortcut" syntax, which facilitates invariance testing with ordinal data (Muthén, 2013).

The PSS is a widely used psychological instrument and its psychometric properties have been previously studied with different populations in various geographic settings (for a review, see Lee, 2012). Lavoie and Douglas (2012) examined measurement invariance of the PSS using WLSMV and polychoric correlations across gender and mental health status of psychiatric patients in the United States. Following recommended procedures (the 3-step approach described below), Lavoie and Douglas found the PSS had configural, metric, and scalar invariance across gender *within* clinical and community groups, but only configural invariance across mental health status.

For simplicity, we describe invariance testing with two groups, but multiple group analyses can involve many groups. In our example, we used Mplus 7.2 (Muthén & Muthén, 1998–2014) to evaluate invariance of the PSS across physical health status. Slightly more than 10% (12.5%, n = 239) of the sample reported having a chronic illness or disability (referred to below as the YES group); 85.5% did not have a chronic condition (n = 1,673, referred to as the NO group). We chose the PSS because this scale has been used widely, and previous research suggested the PSS was likely to present the kinds of challenges social work researchers encounter in their own invariance tests.

Analysis and Evaluation Specification

In Mplus (Muthén & Muthén, 1998–2014), we specified that our items were ordinal by using the variable option CATEGORICAL ARE. We specified WLSMV as the estimator (Beauducel & Herzberg, 2006; Flora & Curran, 2004), although Mplus will use WLSMV by default when categorical variables are specified. We evaluated model fit using the comparative fit index (CFI), Tucker-Lewis index (TLI), and root mean square error of approximation (RMSEA). As indicators of good fit, we recommend using CFI and TLI cutoffs of .95 (or higher) and RMSEA point estimate and upper confidence interval of .06 or lower (West, Taylor, & Wu,

2012). However, we proceeded through the following demonstration of invariance testing even though adequate fit was not obtained at any step. In "real" tests of invariance, researchers would not proceed to scalar testing if the configural or metric models did not have acceptable fit.

The 4-Step Approach. We demonstrate the four steps from Figure 3, in which factor loadings are tested separately from thresholds; however, we also describe the steps and results from the rival approach, in which loadings and thresholds are constrained simultaneously. Steps 2 through 4 correspond to tests of configural, metric, and scalar invariance, respectively.

Step 1: Baseline model for each group. As a first step, we identified the baseline or best-fitting model for each group. Based on prior empirical work on the PSS (e.g., Cohen & Williamson, 1988; Leung, Lam, & Chan, 2010; Reis, Hino, & Rodriguez-Añez, 2010) and our own exploratory factor analysis, we tested a two-factor structure. Per the Mplus default, the first indicator of each factor ("top thing" for Positive and "control" for Negative) was fixed at 1.0 for model identification. Model fit for both groups was inadequate (for the YES disability group: $\chi^2(34) = 111.287$, $p = .0000$; RMSEA = .098(.078, .118); CFI = .883; TLI = .845; for the NO disability group: $\chi^2(34) = 667.919$, $p = .0000$; RMSEA = .106(.099, .113); CFI = .883; TLI = .845).

Modification indices indicated that allowing two pairs of residual covariances in each group to covary would improve fit. Therefore, error terms for "top thing" and "irritate" were allowed to covary in both groups; the errors for "control" and "upset" were allowed to covary in the NO group, and error terms for "confident" and "effective" were allowed to covary in the YES model. (In a "real" analysis, the addition of these error covariances would require theoretical justification.) Fit improved but still failed to meet our prespecified criteria (for the YES group: $\chi^2(32) = 88.426$, $p = .0000$; RMSEA = .086(.065, .107); CFI = .914; TLI = .880; for the NO group: $\chi^2(32) = 316.431$, $p = .0000$; RMSEA = .073(.066, .080); CFI = .948; TLI = .926). However, consistent with published studies (Byrne et al., 1989; Raykov et al., 2012), we proceeded to tests of the configural model with our marginally adequate baseline models.

Step 2: Configural invariance. As illustrated in Figure 3, the group-specific baseline models obtained in Step 1 were entered into a multiple group analysis in Step 2 to test for configural invariance. In Mplus, the configural test is specified with residual variances fixed to 1.0, so all thresholds can be freely estimated across groups (Muthén & Muthén, 1998–2012). Fit statistics were as follows: $\chi^2(64) = 394.460$, $p = .0000$; RMSEA = .073(.067, .081); CFI = .944; TLI = .921. None of the indices met our prespecified criteria. In a real analysis, researchers would conclude that either the configural model is different for the two groups, or that the groups share an inadequate configural model. In the case of different models, analyses using the latent construct would need to be conducted separately for the

two groups. In the case of inadequate fit, the validity of the scale for use in research or practice would be in question.

For the purpose of demonstrating the next steps in invariance testing, we proceeded to test for metric invariance despite the finding of inadequate configural fit. For the rest of the example, we focus on invariance test statistics and not overall fit; in real analyses, the lack of fit would lead to the conclusion that the scale is inadequate for use in research and practice.

Step 3: Metric invariance. In Step 3, the fit of the model with all factor loadings constrained across the two health status groups was compared to the fit of the configural model. The change in χ^2 per change in *df* was nonsignificant [$\Delta\chi^2(8) = 7.040$, $p = 0.5323$], indicating that loadings for the two groups were statistically equivalent. With this evidence of metric invariance, it was appropriate to retain the constrained loadings and proceed to a test of scalar invariance. On its own, metric invariance does not provide adequate justification for using the Positive and Negative factors in practice or research with data from the two groups combined.

Step 4: Scalar invariance. The fit of a model with factor loadings and all thresholds constrained to be equal across the two health status groups was compared to the fit of the final metric model. The change in χ^2 was statistically significant [$\Delta\chi^2(28) = 94.465$, $p = 0.0000$], indicating one or more thresholds was noninvariant across groups. As described in the box to the right of Step 4 in Figure 3, we began the search for noninvariant thresholds by backing up to the metric model and then constraining all thresholds on one factor at a time. We hoped to find invariance across all thresholds on Positive, and avoid having to test thresholds one at a time. Note that the decision to start with the Positive factor was arbitrary. We constrained all thresholds on the Positive factor and freed all thresholds on the Negative factor, with the exception of those required to be constrained for identification purposes. Unfortunately, the χ^2 comparison again indicated a significant deterioration in fit between the metric model and the model with thresholds constrained for Positive [$\Delta\chi^2(17) = 47.827$, $p = 0.0001$].

The next step was to look systematically within Positive for one or more noninvariant thresholds causing the significant decrement in fit. We backed up again to the metric model and constrained the third and fourth thresholds on the referent indicator ("top thing"). The first two thresholds were already constrained for identification. Once again, fit deteriorated significantly [$\Delta\chi^2(2) = 23.820$, $p = 0.0000$]. The finding suggested that one or both of the tested thresholds was noninvariant. We proceeded to test each threshold separately, comparing χ^2 of the model with the threshold constrained to the χ^2 of the metric model. Each time fit became significantly worse [$\Delta\chi^2(1) = 17.151$, $p = 0.0000$ for the third threshold, and $\Delta\chi^2(1) = 17.435$, $p = 0.0000$ for the fourth threshold], so we concluded both thresholds were noninvariant, and therefore, left them free.

We proceeded to test the non constrained thresholds for "success" (the second, third, and fourth thresholds), the next indicator of Positive. Fit declined significantly when we constrained all three thresholds at once [$\Delta\chi^2(3) = 31.824$, $p = .0000$], so we proceeded to test each one separately. All three proved to be noninvariant [$\Delta\chi^2(1) = 19.439$, $p = .0000$ for the second threshold, $\Delta\chi^2(1) = 20.229$, $p = .0000$ for the third, and $\Delta\chi^2(1) = 19.743$, $p = .0000$ for the fourth]. Fit also declined significantly when we constrained all three non constrained thresholds of the next indicator of Positive, "effective" [$\Delta\chi^2(3) = 28.775$, $p = .0000$] so we proceeded to test each one separately. Again, each individual threshold was noninvariant [$\Delta\chi^2(1) = 9.295$, $p = 0.0023$ for the second threshold, $\Delta\chi^2(1) = 19.350$, $p = 0.0000$ for the third, and $\Delta\chi^2(1) = 17.278$, $p = 0.0000$ for the fourth].

The eight noninvariant thresholds that we found so far represented 26.7% of the 30 loading and thresholds parameters for Positive, exceeding the cutoff suggested by Dimitrov (2010). Therefore, we had to conclude that the Positive factor of the PSS is not invariant across health status groups. Corresponding tests for threshold invariance yielded the same results for the Negative factor.

Interpretation. Overall, findings from our invariance tests of the PSS indicated that it had configural and metric invariance, but not scalar invariance across groups with and without chronic illness or disability. The configural invariance finding suggests that the two dimensional structure of the PSS applies to both groups. The metric invariance finding suggests that individual items have similar weights and are equally salient to the construct of perceived stress for both groups. However, the finding of scalar noninvariance reveals that similar true levels of the Positive and Negative dimensions of the PSS may correspond to different response choices across groups on indicators of the latent variable. Conversely, in some cases, different true levels of perceived stress might correspond to the same observed score across the two groups.

Configural and metric invariance are not considered adequate for interpreting scale scores the same across groups. In the absence of scalar invariance, mean scores on the PSS factors cannot be validly compared across groups in research or practice, nor should composites or latent variables representing PSS scores be used in hypothesis testing across the two groups or in practice.

The 3-Step Approach. We repeated invariance testing of the PSS across the two groups with and without chronic health issues using the 3-step approach that omits separate tests of factor loadings (metric invariance). Supporters of this approach claim that factor loadings and thresholds should be freed or constrained together. Steps 1 and 2 in Figure 3 were the same as described above. Step 3 in this approach involved comparing the fit of the final configural model to the scalar model, that is, the model with both factor loadings and thresholds constrained to be equal.

Fit of the scalar model was significantly worse than fit of the final configural model [$\Delta\chi^2(36) = 96.608$, $p = .0000$], indicating that one or more indicators had noninvariant parameters. We next tested a model with only the loadings and thresholds of indicators of Positive constrained across groups. Again, fit was significantly worse than the final configural model [$\Delta\chi^2(22) = 51.034$, $p = .0004$], so we began to test the set of loading and threshold parameters for one indicator of Positive at a time. Starting with "top thing" we constrained the third and fourth thresholds ("top thing" loading was already fixed to 1.0 in both groups, and its first two thresholds were constrained for identification purposes). All other loadings and thresholds that did not have to remain constrained for indicators of Positive were allowed to vary across groups. Compared to the final configural model, the two constraints did not significantly reduce fit [$\Delta\chi^2(2) = 3.750$, $p = 0.1534$], suggesting the thresholds were invariant and could ultimately be constrained to be equal. For the moment, we re-freed the two thresholds so we could proceed to test the loading for "success" and its three non constrained thresholds. The four constraints did not significantly reduce fit [$\Delta\chi^2(4) = 6.918$, $p = .1403$], meaning the parameters were invariant. We temporarily freed the parameters and tested, in turn, the loading and threshold set for each of the next four indicators of Positive. Fit did not deteriorate significantly when constraints were added individually for "effective" [$\Delta\chi^2(4) = 3.464$, $p = .4833$], for "confident" [$\Delta\chi^2(4) = 4.818$, $p = .3065$], for "things go" [$\Delta\chi^2(4) = 6.798$, $p = .1470$], or for "irritate" [$\Delta\chi^2(4) = 3.947$, $p = .4133$].

When all loadings and thresholds of indicators of Positive were constrained simultaneously, we expected to find one or more indicators that caused the significant decrease in fit that was observed. However, because no one indicator caused a significant reduction in χ^2 by itself, we needed to find the combination of indicators that explained the noninvariance. We have not seen this situation discussed in the literature. We ranked the six indicators in terms of the $\Delta\chi^2$ observed when loadings and thresholds were constrained. We then proceeded to constrain loading and threshold sets in order, starting with the items whose constraints had the least effect on χ^2. For example, the $\Delta\chi^2$ obtained when constraining loadings and thresholds for "effective" was the smallest (3.464); and the $\Delta\chi^2$ for "top thing" was the next smallest (3.750). We constrained the loadings and thresholds for these two indicators and examined the $\Delta\chi^2$ relative to the configural model. The change in fit was nonsignificant [$\Delta\chi^2(6) = 3.620$, $p = .7280$]. We kept these constraints and added constraints to the loading and thresholds for "irritate," the indicator with the next smallest $\Delta\chi^2$. Again, $\Delta\chi^2$ was nonsignificant [$\Delta\chi^2(10) = 7.688$, $p = .6593$], suggesting the parameters of the three indicators were invariant. Constraining the loading and thresholds for "confident" in addition to those for the previous three indicators led to the same conclusion [$\Delta\chi^2(14) = 19.258$, $p = .1553$]. However, when we added constraints on the parameters associated with "success," the decrease in χ^2 relative to the configural

model was significant [$\Delta\chi^2(18) = 31.473$, $p = .0254$]. Therefore, we concluded that parameters of "success" and "things go" (the indicator with the highest $\Delta\chi^2$ in the individual tests) were noninvariant. Eight parameters were associated with the two noninvariant indicators (two loadings and six thresholds). A total of 30 parameters were associated with indicators of Positive. With more than a quarter of the factor's parameters noninvariant (8/30 = 27%), we had to conclude that the Positive dimension of the PSS was noninvariant across health status groups.

Repeating the above steps for the Negative factor, we encountered the same pattern—constraining all of the indicators' loadings and thresholds at the same time led to a significant decrement in fit [$\Delta\chi^2(14) = 52.641$, $p < .0001$], but no single indicator was noninvariant when testing individually. Adding loading and threshold constraints indicator by indicator, starting with the one with the smallest individual $\Delta\chi^2$, led to the conclusion that the Negative dimension of the PSS was also noninvariant.

Comparison of Conclusions Drawn from the Two Approaches

Both approaches to invariance testing of the PSS across health status groups led to the same conclusion: the Positive and Negative dimensions of perceived stress were noninvariant. The dimensions did not pass the scalar invariance test with either approach. However, our example does not allow us to claim that the two approaches will always lead to the same conclusion about a scale's invariance. Social workers will need to make a decision about which approach they prefer based on the discussion above or their own reading of the sources cited here and elsewhere. Using the 3-step approach, with which multiple parameters associated with one item are freed even if only one is noninvariant, is a more conservative approach. Specifically, with the 3-step approach, measures are more likely to be found noninvariant across groups because Dimitrov's (2010) 20% cutoff for the ratio of noninvariant parameters to the total number of loadings and thresholds associated with a factor will be reached more quickly.

Conclusion

Our review of social work literature published from 2002 to 2014 indicates that the appropriate treatment of ordinal data in measurement invariance tests is rare. As a guide for social work researchers, we have provided background information and recommendations based on our review of the literature. Through an example, we have illustrated two competing approaches: a 4-step approach, in which factor loadings and thresholds are tested and constrained separately; and a 3-step approach, in which loadings and thresholds are tested and constrained in tandem. Both approaches led to the same conclusion of noninvariance for the PSS. The example illustrated authentic decision points that may be encountered

by social work researchers for which the literature does not provide practical guidance. We demonstrated one systematic response to an unexpected finding—noninvariance for a set of parameters that could not be attributed to any one parameter. In their own research, social workers are likely to encounter other such situations demanding analogous systematic responses.

The PSS provides an example of a scale that is commonly used in social work research and practice, yet failed to pass even the preliminary step for invariance testing in the current sample—the development of a baseline, group-specific model with adequate fit. The best-fitting baseline models for the two groups examined failed to meet prespecified fit criteria, calling into question the validity of the scale for individuals with and without chronic illness or disability. In addition, the two dimensions of the scale were noninvariant, suggesting scores should not be compared across the two groups. Careful examination of new and existing social work scales is likely to reveal fit or noninvariance problems across some groups. For many scales, the problems might be minimal and ignorable, or fixable. For others, the conclusions of invariance tests will indicate the need to identify or develop new measures, or interpret scale scores differently for members of different groups. Most important, appropriate analysis of scale data will help social workers avoid using scales from which they could draw erroneous and potentially harmful research and practice conclusions.

Author Notes

Natasha K. Bowen is a professor in the School of Social Work at the University of North Carolina at Chapel Hill.

Rainier D. Masa is a doctoral candidate, research assistant, and adjunct instructor in the School of Social Work at The University of North Carolina at Chapel Hill: rmasa@email.unc.edu

Correspondence regarding this article should be sent to Dr. Natasha Bowen, UNC School of Social Work, 325 Pittsboro St CB #3550 Chapel Hill, NC 27599-3550, or via e-mail to nbowen@email.unc.edu

References

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

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Bovaird, J. A., & Koziol, N., A. (2012). Measurement models for ordered-categorical indicators. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 495—511). New York, NY: Guilford Press.

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464—504. http://dx.doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (1998). Cross-cultural comparisons using noninvariant measurement items. *Applied Behavioral Science Review, 6*, 93—110. http://dx.doi.org/10.1016/S1068-8595(99)80006-3

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1—27. http://dx.doi.org/ 10.1177/014920639902500101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233—255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*, 385—396. http://dx.doi.org/10.2307/2136404

Cohen, S., & Williamson, G. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The social psychology of health: Claremont symposium on applied social psychology* (pp. 31—67). Newbury Park: CA: Sage.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*, 121—149. http://dx.doi.org/10.1177/0748175610373459

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

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13*, 378—402. http://dx.doi.org/10.1207/s15328007sem1303_3

Godfrey, E. B., Osher, D., Williams, L. D., Wolf, S., Berg, J., Torrente, C., . . . Aber, J. L. (2012). Cross-national measurement of school learning environments: Creating indicators for evaluating UNICEF's child friendly schools initiative. *Children and Youth Services Review, 34*, 546—557. http://dx.doi.org/10.1016/j.childyouth.2011.10.015

Granillo, M. T. (2012). Structure and function of the Patient Health Questionnaire-9 among Latina and non-Latina White female college students. *Journal of the Society for Social Work and Research, 3*, 80—93. http://dx.doi.org/10.5243/jsswr.2012.6

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling, 16*, 642—657. http://dx.doi.org/10.1080/10705510903206014

Jöreskog, K. G. (2005). *Structural equation modeling with ordinal variables using LISREL.* Lincolnwood, IL: Scientific Software International.

Lavoie, J. A. A., & Douglas, K. S. (2012). The perceived stress scale: Evaluating configural, metric, and scalar invariance across mental health status and gender. *Journal of Psychopathology and Behavioral Assessment, 34*, 48—57. http://dx.doi.org/10.1007/s10862-011-9266-1

Lee, E. H. (2012). Review of the psychometric evidence of the perceived stress scale. *Asian Nursing Research, 6*, 121—127. http://dx.doi.org/10.1016/j.anr.2012.08.004

Leung, D. Y., Lam, T., & Chan, S. S. (2010). Three versions of perceived stress scale: Validation in a sample of Chinese cardiac patients who smoke. *BioMed Central Public Health, 10*, 513. http://dx.doi.org/10.1186/1471-2458-10-513

Lubke, G. H., & Muthén, B. O. (2004). Applying multipgroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*, 514—534. http://dx.doi.org/10.1207/s15328007sem1104_2

McLarnon, M. J. W., & Carswell, J. J. (2013). The personality differentiation by intelligence hypothesis: A measurement invariance investigation. *Personality and Individual Difference, 54*, 557–561. http://dx.doi.org/10.1016/j.paid.2012.10.029

Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). New York, NY: Guilford Press.

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

Muthén, B. O. (2013). *Version 7.1 Mplus Language Addendum.pdf*. Retrieved from http://www.statmodel.com/download/Version7.1xLanguage.pdf

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus.* Mplus Web Note #4.

Muthén, L. K., & Muthén, B. O. (1998–2014). *Mplus, v.7.2.* Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Raykov, T., Marcoulides, G. A., & Li, C. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement, 72*, 954–974. http://dx.doi.org/10.1177/0013164412441607

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement, 73*, 713–727. http://dx.doi.org/10.1177/0013164412451978

Reis, R. S., Hino, A. A. F., & Rodriguez-Añez, C. R. (2010). Perceived stress scale: Reliability and validity study in Brazil. *Journal of Health Psychology, 15*, 107–114. http://dx.doi.org/10.1177/1359105309346343

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*, 347–363. http://dx.doi.org/10.1177/0734282911406661

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling, 21*, 167–180. http://dx.doi.org/10.1080/10705511.2014.882658

Sellers, S. L., Mathieson, S. G., Smith, T., & Perry, R. (2006). Perceptions of professional social work journals: Findings from a national survey. *Journal of Social Work Education, 42*, 139–160. http://dx.doi.org/10.5175/JSWE.2006.200303095

Silver Wolf, D. A. P., Dulmus, C. N., Maguin, E., & Fava, N. (2014). Refining the evidence-based practice attitude scale: An alternative confirmatory factor analysis. *Social Work Research, 38*, 47–58. http://dx.doi.org/10.1093/swr/svu006

Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods, 7*, 210–227. http://dx.doi.org/10.1037/1082-989X.7.2.210

Thyer, B. A. (2008). *Preparing research articles*. New York, NY: Oxford University Press.

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

Webber, K. C. (2014). *School engagement of rural early adolescents: Examining the role of academic relevance and optimism across racial/ethnic groups*. Unpublished doctoral dissertation. The University of North Carolina at Chapel Hill.

Wegmann, K. M. (2014). *A mixed-methods exploration of stereotype threat in middle childhood.* Unpublished doctoral dissertation. The University of North Carolina at Chapel Hill.

West, S. G., Taylor A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford Press.

Manuscript submitted: September 29, 2014
Revision submitted: December 17, 2014
Accepted: January 13, 2014
Electronically published: April 6, 2015