Psychological Assessment
2015, Vol. 27, No. 1, 90–101

© 2014 American Psychological Association
1040-3590/15/$12.00 http://dx.doi.org/10.1037/a0038100

Psychometric Analysis of the Ten-Item Perceived Stress Scale

John M. Taylor
University of Missouri—Columbia

Although the 10-item Perceived Stress Scale (PSS-10) is a popular measure, a review of the literature reveals 3 significant gaps: (a) There is some debate as to whether a 1- or a 2-factor model best describes the relationships among the PSS-10 items, (b) little information is available on the performance of the items on the scale, and (c) it is unclear whether PSS-10 scores are subject to gender bias. These gaps were addressed in this study using a sample of 1,236 adults from the National Survey of Midlife Development in the United States II. Based on self-identification, participants were 56.31% female, 77% White, 17.31% Black and/or African American, and the average age was 54.48 years ($SD = 11.69$). Findings from an ordinal confirmatory factor analysis suggested the relationships among the items are best described by an oblique 2-factor model. Item analysis using the graded response model provided no evidence of item misfit and indicated both subscales have a wide estimation range. Although t tests revealed a significant difference between the means of males and females on the Perceived Helplessness Subscale ($t = 4.001$, $df = 1234$, $p < .001$), measurement invariance tests suggest that PSS-10 scores may not be substantially affected by gender bias. Overall, the findings suggest that inferences made using PSS-10 scores are valid. However, this study calls into question inferences where the multidimensionality of the PSS-10 is ignored.

Keywords: differential item functioning, graded response model, item response theory, confirmatory factor analysis, perceived stress

The Perceived Stress Scale (PSS) is a self-report measure intended to capture the degree to which persons perceive situations in their life as excessively stressful relative to their ability to cope (Cohen, Kamarck, & Mermelstein, 1983). The PSS has emerged as the most popular measure of perceived stress (Karam et al., 2012). It has been translated into 25 different languages (Cohen, 2013), validated on diverse samples (Mitchell, Crane, & Kim, 2008), and used across a broad range of fields to answer empirical questions and guide clinical practice (e.g., Roberti, Harrington, & Storch, 2006). Despite this popularity, interest in the PSS's psychometric properties is a relatively recent occurrence and little is known about the measure's psychometric capabilities in the extant literature. The purpose of this article is to address important gaps in the psychometric literature of the 10-item version of the PSS (PSS-10).

The PSS was developed to serve as a global, subjective measure of perceived stress (Cohen et al., 1983). Cohen et al. (1983) noted at the time that stress measures tended to assess stress objectively (e.g., frequency of stressful stimuli) and emphasize specific events (e.g., job loss) while ignoring the cognitive appraisal process individuals engage in when they encounter stressful stimuli. Cohen et al. (1983) argued that ignoring the appraisal process was limiting and developed the PSS using Lazarus's original transactional stress model to address the problem—a model that characterizes stress in terms of the interchange between the appraisal of the stressor (e.g., severity) and one's perceived ability to cope (Shewchuk, Elliott, MacNair-Semands, & Harkins, 1999). Although the PSS was developed to primarily serve researchers' interests, it has since been used in clinical settings (Cohen et al., 1983; Mitchell et al., 2008). Cohen and Williamson (1988) resisted the idea of the PSS as being a diagnostic measure, but Cohen et al. (1983) claimed the instrument likely taps prodromal stages of psychiatric disorders. Thus, the PSS has since been recommended as a means of identifying individuals at risk for worsening conditions, a tool to aid clinicians in treatment planning, and a means of tracking a client's response to an intervention (Roberti et al., 2006).

To date, there are three standard versions of the PSS: the original 14-item form (PSS-14), the PSS-10, and a four-item form (PSS-4; Cohen et al., 1983). Cohen et al. (1983) reported that scores on the PSS-14 exhibited good consistency (e.g., Cronbach's alpha was .86 among participants in a smoking-cessation intervention) and moderate predictive and concurrent validity. However, using exploratory factor analysis (EFA) in a subsequent study, Cohen and Williamson (1988) identified four poorly performing items and dropped them from the PSS-14 giving rise to the more commonly used PSS-10. In addition, Cohen and Williamson (1988) further shortened the measure to the PSS-4 for situations where measurements need to be taken quickly. Cohen and Williamson (1988) reported that scores on both the PSS-10 and PSS-4 demonstrated moderate convergent validity, but scores from the PSS-4 exhibited relatively low reliability ($\alpha = .60$) compared to scores produced by the PSS-10 ($\alpha = .78$). As a result, they suggested that the PSS-10 is the best form of the PSS and recommended the PSS-10 be used in future research.

In subsequent studies PSS-10 scores have continued to exhibit good measurement properties consistent with the original findings

This article was published Online First October 27, 2014.

John M. Taylor, Department of Educational, School, and Counseling Psychology, University of Missouri—Columbia.

Correspondence concerning this article should be addressed to John M. Taylor, who is now at School of Nursing, Saint Louis University, St. Louis, MO 63103. E-mail: johntaylor@slu.edu

of Cohen and Williamson (1988). For example, Karam et al. (2012) reported that PSS-10 scores exhibited good reliability in a sample of pregnant women taking an antidepressant (α = .90), including better reliability than scores produced by the PSS-4 (α = .79). Mitchell et al. (2008) found PSS-10 scores to exhibit good convergent validity and reported some evidence of concurrent validity. Similar reliability and validity findings have been observed across cultures as well (e.g., Ramírez & Hernanzez, 2007; Reis, Hino, & Rodriguez-Anez, 2010; Remor, 2006). These studies highlight, however, that the psychometric literature on the PSS-10 has been limited to aggregate-level methods.

While aggregate-level methods are useful, they provide limited information at the item level. Most previous analyses of the PSS assume the items estimate perceived stress across the latent continuum equally well and do not directly assess the fit of the items to a model (Embretson & Reise, 2000). Thus, there is a need to complement classical psychometric methods with modern methods. Item response theory (IRT) is well-equipped to assess model-data fit of individual items and to identify the estimation capacity of each item at different points along the latent continuum (de Ayala, 2009). Yet IRT analysis has not been used to assess the performance of scores on the PSS-10 at the item level. One exception is that Sharp, Kimmel, Kee, Saltoun, and Chang (2007) reported using IRT to study the PSS; however, none of the IRT results were provided (e.g., item location) as their interest in IRT was limited to tests for racial bias.

Gender Bias

The study by Sharp et al. (2007) introduces the fact that bias is a significant issue for the PSS in general and the PSS-10 specifically. Bias is a test characteristic involving the degree of disparity between respondents' true score on a latent variable and manifest score on the latent indicator (e.g., total test score, subscale score, item score; Sass, 2011). When bias is present, parameters are likely subject to systematic under- or overestimation (de Ayala, 2009). Researchers in the last decade have been particularly interested in bias that varies as a function of sample characteristics (i.e., measurement noninvariance; Milfont & Fischer, 2010; Sass, 2011). The interest is understandable given that noninvariant measures can undermine the validity of inferences, particularly inferences involving group comparisons (Zumbo, 1999). Namely, when variance in a measure is a function of the underlying latent variable and an unintended secondary factor such as group membership (i.e., bias), differences observed between groups are not necessarily indicative of differences on the latent variable (Sass, 2011).

The PSS has been studied in samples with diverse characteristics (e.g., pregnant females taking an antidepressant) often for the purpose of identifying whether the psychometric properties of the PSS (e.g., factor structure) remain invariant under changing conditions (e.g., Karam et al., 2012; Ramírez & Hernanzez, 2007). Such studies demonstrate the consistency of the PSS's properties across diverse samples, but most have not directly tested for noninvariance. Direct tests for invariance typically take the form of one or more differential item functioning (DIF) methods or multigroup confirmatory factor analysis (MCFA; Stark, Chernyshenko, & Drasgow, 2006). Although positive findings from DIF tests or MCFA are only cause for further study, if measurement invariance is not supported through one of these approaches then bias can be suspected in the latent indicator (Zumbo, 1999). In most PSS studies that have employed direct tests for noninvariance evidence of bias has been found (e.g., Sharp et al., 2007).

Of the possible sources of bias, gender bias has been the most salient for the PSS. Cohen and Williamson (1988) found a statistically significant difference between males' and females' total scores on the PSS-10 with females reporting higher levels of overall perceived stress than males, on average. In studies dividing the PSS-10 and 14 into two subscales, females commonly report higher scores among the negatively phrased items (i.e., Perceived Helplessness Subscale [PHS]) compared to males but not among the remaining positively phrased items (i.e., Perceived Self-Efficacy Subscale [PSES]; e.g., Hewitt, Flett, & Mosher, 1992; Roberti et al., 2006). Several explanations have been put forth to explain mean gender differences, including that the PSS is gender biased (Lavoie & Douglas, 2012).

However, the few studies that have directly examined the possibility of gender bias in the PSS have not produced consistent findings. For example, Gitchel, Rosessler, and Turner (2011) used the poly-SIBTEST to conduct a DIF test on six negatively phrased items from the PSS-14. Four of the items were found to exhibit DIF in an adult sample diagnosed with multiple sclerosis (Gitchel et al., 2011). In contrast, Lavoie and Douglas (2012) used confirmatory factor analysis (CFA) to test for noninvariance in a recently discharged adult psychiatric sample and found none in the PSS-14. Lavoie and Douglas (2012) speculated that the inconsistency between their findings and Gitchel et al.'s (2011) could be attributed to the different types of methods used or the different samples used. The issue of gender bias remains unclear.

Factor Structure

Prior to tackling the invariance issues of the PSS-10, or analysis of the items, the dimensionality of the PSS-10 needs to be firmly established (de Ayala, 2009). Although results from an EFA suggested the intercorrelations of both the PSS-14 and PSS-10 are explained by two latent factors, Cohen and Williamson (1988) dismissed the second factor as "irrelevant" (p. 43). Since then, nearly all studies have found results similar to that of Cohen and Williamson, with the negatively phrased items loading onto the perceived helplessness factor and the positively phrased items loading onto the perceived self-efficacy factor (e.g., Roberti et al., 2006). Unlike Cohen and Williamson, other researchers do not reject the second factor as irrelevant. For example, Hewitt et al. (1992) studied the factor structure of the PSS-14 on a psychiatric sample and found that the two factors identified through EFA made distinct contributions in a subsequent regression analysis. Both PHS and PSES predicted depression in women but only the PHS predicted depression in men (Hewitt et al., 1992). The distinct predictive quality of the subscales suggests the two-factor solution is relevant.

But studies such as Hewitt et al. (1992) have methodological limitations that are important for understanding whether this is a unidimensional measure. First, there is some concern that the manifestation of two latent factors in EFA is an artifact of psychiatric samples (Lavoie & Douglas, 2012). Second, with few exceptions, EFA has been the method used to determine the number of factors underlying the PSS. Although a powerful tool, EFA is not without problems, such as the criteria researchers use to select the number of underlying factors (Ledesma & Valero-Mora, 2007). For example, Hewitt et al. (1992) used a scree plot to determine the number of factors that underlie the PSS-14. But scree plots have been criticized for being too subjective, and researchers who use them often overestimate the number of factors underlying a measure (Ledesma & Valero-Mora, 2007). It is possible that studies such as Hewitt et al. (1992) overestimate the number of factors underlying the PSS-10.

Still, the cumulative evidence from the literature favors a two-factor model, including with nonclinical samples. Multiple criteria have been applied across studies to determine the number of factors, and the results have been highly consistent regardless of the method used or sample studied. For example, Roberti et al. (2006) found that both the Kaiser criterion and screen plot suggested two factors underlie the PSS-10 in a sample of university students recruited from various introductory courses. The results were supported by a CFA that suggested the two-factor solution exhibits good fit (Roberti et al., 2006).

Nevertheless, a comparison of competing hypothesized models has not been fully addressed in the literature. Two studies provide limited evidence. Barbosa-Leiker et al. (2013) studied the PSS-10 using CFA and found the unidimensional model tended to fit the data poorly while the two-factor model consistently fit the data well. Leung, Lam, and Chan (2010) reported similar findings from a CFA study of the Chinese version of the PSS-10. Unfortunately, in both studies the ordered categories of the PSS-10 items were treated as continuous, a practice generally discouraged since it can undermine the validity of inferences (Flora & Curran, 2004). Simulation studies have demonstrated that normal-theory CFA methods tend to underperform ordinal CFA methods when analyzing ordered categories (e.g., exaggerates the misfit of models) and it has been observed in the literature that choice of ordinal CFA methods over that of normal-theory CFA methods can influence outcomes (e.g., Holgado-Tello, Carrasco-Ortiz, Gándara, & Moscoso, 2009; Holgado-Tello, Chacón-Mascoso, Barbero-García, & Vila-Abad, 2010). As a result, it remains unclear whether the one-factor model of the PSS-10 shows good fit and whether a multidimensional model is an improvement over a unidimensional model that justifies rejection of the more parsimonious unidimensional model.

Present Study

The discussion above highlights three significant gaps in the psychometric literature of the PSS-10. First, there is some debate as to whether a one-factor or a two-factor model best describes the underlying factor structure of the PSS-10. Second, there is little information available on the performance of the individual items that make up the PSS-10. In particular, the estimation capacity of each item at different points along the latent continuum is unknown, and it is also unknown whether one or more of the items exhibit poor model-data fit. Finally, it is unclear whether scores on the PSS-10 are subject to gender bias. The concern is that the tendency of female respondents' to report higher levels of global perceived stress, or perceived helplessness, than males on the PSS-10 may not represent genuine differences on the latent variable(s; Lavoie & Douglas, 2012). Thus, the aim of this study was threefold: (a) clarify the dimensionality of the PSS-10, (b) assess the performance of the items, and (c) test for DIF.

Method

Participants

The present analyses uses data from the National Survey of Midlife Development in the United States II (MIDUS II; Ryff, Seeman, & Weinstein, 2004–2009). According to Brim, Ryff, and Kessler (2004), the MIDUS data consists of English-speaking adults available through random digit dialing from within the 48 contiguous United States. The purpose of MIDUS was to conduct a national longitudinal study of the links between "psychological and social factors" and an array of health outcomes of middle-aged and older adults (Love, Seeman, Weinstein, & Ryff, 2010, p. 1059). The first wave (MIDUS I) was initiated in 1995 and was composed of 7,108 adults between the ages of 25 and 74 years (Love et al., 2010; Morozink, Friedman, Coe, & Ryff, 2010). The second wave (MIDUS II) was initiated in 2004 and was composed of five distinct projects (Love et al., 2010).

Analyses in this study are based upon 1,236 adults from the Biomarker project, which is a subset of MIDUS II participants (Love et al., 2010). Love et al. (2010) reported that the Biomarker project added a broad array of biological data to the MIDUS data. To be eligible participants had to be healthy enough to travel to one of three clinics at University of California, Los Angeles; University of Wisconsin; or Georgetown University for 2 days of assessments (Love et al., 2010). Ages reported in the present sample ranged from 34 to 83 years ($M = 54.48$, $SD = 11.69$). Five-hundred forty participants were males, and 696 were females. Approximately 77% identified themselves as White; 17.31% identified themselves as Black and/or African American; and 2.5% identified themselves as Native American or Aleutian Islander/Eskimo, Asian or Pacific Islander, multiracial, or other. According to Morozink et al. (2010), participants in the Biomarker project are similar to participants in the MIDUS II except that participants in the Biomarker project reported higher levels of education.

Measure

The PSS-10 is a self-report measure consisting of 10 items purported to measure "how unpredictable, uncontrollable, and overloaded respondents find their lives" (Cohen & Williamson, 1988, p. 34). According to Cohen and Williamson (1988), the instrument was designed for use in community samples and assumes respondents have at least a middle school education. Respondents complete the PSS-10 on a Likert-type scale with response categories ranging from 1 (*Never*) to 5 (*Very often*) and total scores are tallied by reverse-scoring Items 4, 5, 7, and 8 and then summing across all 10 items (Cohen et al., 1983; Cohen & Williamson, 1988). Consistent with previous studies, reliability of the overall measure in this sample was .84 while the reliabilities of the PHS and PSES were .86 and .82, respectively (e.g., Roberti et al., 2006).

Data Analysis

In order to address the first goal three separate factor models were tested using ordinal CFA. The first model specified was a unidimensional congeneric model with all 10 items loaded onto one latent factor (i.e., original hypothesized factor structure; Cohen

& Williamson, 1988). The next two models were an orthogonal two-factor solution and an oblique two-factor solution. In both cases the six negatively phrased items were loaded onto the perceived helplessness factor and the remaining four positively phrased items[1] were loaded onto the perceived self-efficacy factor (Hewitt et al., 1992; Roberti et al., 2006). All models were estimated by analyzing an asymptotic polychoric covariance matrix using robust unweighted least squares estimation in LISREL 8.80 (Jöreskog & Sörbom, 2006). Model-data fit was assessed using the comparative fit index (CFI), Tucker–Lewis index (TLI), and root-mean-square error of approximation (RMSEA), indices that are commonly reported in CFA studies with ordinal indicators. Although the literature provides little guidance regarding assessment of model fit with ordinal indicators, prior studies conducting similar analyses have typically considered CFI and TLI values greater than 0.95 as evidence of good model-data fit along with RMSEA values less than 0.06 (e.g., Lavoie & Douglas, 2012; Sass, 2011). These same guidelines were employed in this study.

Typically, the graded response model (GRM) and rating scale model (RSM) are used to estimate IRT parameters for Likert-type items (de Ayala, 2009). However, there is some evidence that the RSM performs poorly when tests are 20 items or less (Wang & Chen, 2005). In contrast, the GRM estimates item parameters well in short tests since estimation bias appears to be influenced by sample size rather than test length (Kieftenbeld & Natesan, 2012). Therefore, the GRM was implemented in the present study to address the second goal. Samejima's (1969) GRM is defined as (de Ayala, 2009, p. 219):

$$P^*_{x_j}(\theta) = \frac{e^{\alpha_j(\theta-\delta_{x_j})}}{1 + e^{\alpha_j(\theta-\delta_{x_j})}}, \tag{1}$$

where the probability of crossing threshold *x* or higher, on item j, is determined by the person's location θ, the threshold's location parameter δ_{x_j}, and the discrimination parameter α_j.

IRT parameters were estimated in the R package *ltm* using the unconstrained GRM (Rizopoulos, 2006). However, *ltm* does not provide item-fit indices for the GRM. Instead, item-fit was inspected visually using the program MODFIT (Stark, 2002). In MODFIT an observed category probability curve, or empirical option response function (EMP), is plotted against a predicted category probability curve, or predicted option response function (ORF), for each item's response categories (de Ayala, 2009). If an observed item curve shows significant deviation from the predicted item curve, then the item is suspected of misfit (de Ayala, 2009). Although item-fit can be assessed using χ^2 values provided by MODFIT, χ^2 tests are overpowered when based upon large samples, falsely flagging items as exhibiting misfit (Cheung & Rensvold, 2002). Note that MODFIT requires an instrument's lowest response category to be coded as 0 (Stark, 2002). So the PSS-10's items were recoded for the MODFIT analysis with response categories ranging from 0 (*Never*) to 4 (*Very often*).

As discussed above, DIF methods and MCFA are commonly used frameworks to assess noninvariance of latent indicators and both are used in the present study (Stark et al., 2006). There are few DIF methods available for analyzing short tests. However, since there is some evidence that ordinal logistic regression (OLR) adequately controls type I and type II error rates in such cases, OLR was used to address the third goal (Scott et al., 2009). OLR confers the additional advantage of testing for uniform and nonuniform DIF (Zumbo, 1999). Uniform DIF occurs when the probably of endorsing a response category is higher for one group compared to another group across all levels of the latent variable's continuum (de Ayala, 2009; Zumbo, 1999). Nonuniform DIF is similar to uniform DIF except that the group difference is no longer constant (Zumbo, 1999). Nonuniform DIF is an interaction where the probably of endorsing a response category is higher for one group compared to another group at one end of the latent continuum but lower at the opposite end (de Ayala, 2009). In essence, uniform DIF tests whether item locations, analogous to intercepts in CFA, are invariant and nonuniform DIF tests whether item locations and slopes are invariant (de Ayala, 2009; Stark et al., 2006). Although nonuniform differential test functioning has been investigated in the PSS, to this author's knowledge, nonuniform DIF has not been examined in the PSS before (Lavoie & Douglas, 2012).

Consistent with standard DIF practices, three proportional odds models were specified: The nonuniform DIF model, uniform DIF model, and the base model (Zumbo, 1999). Each model was estimated using the R package *Ordinal* (Christensen, 2011). The nonuniform model in the present study is defined as (Zumbo, 1999):

$$logit\left[P\left(Y \leq j\right)\right] = \alpha_j + b_1X_{total} + b_2X_{gender} + b_3X_{total\ by\ gender}, \tag{2}$$

where the probability of a respondent crossing threshold *j* with intercept α_j is a function of the respondent's total score on the scale (X_{total}), gender (X_{gender}), and the interaction of gender and total score ($X_{total\ by\ gender}$). The uniform model is defined similarly (Zumbo, 1999):

$$logit\left[P\left(Y \leq j\right)\right] = \alpha_j + b_1X_{total} + b_2X_{gender}. \tag{3}$$

The base model is defined as (Zumbo, 1999):

$$logit\left[P\left(Y \leq j\right)\right] = \alpha_j + b_1X_{total}. \tag{4}$$

DIF was tested by examining the ratio of likelihoods between models (ΔG^2)—the ratio between the nonuniform and uniform model to test for nonuniform DIF and the ratio between the uniform and base model to test for uniform DIF (de Ayala, 2009). If ΔG^2, which is χ^2 distributed with one *df*, was statistically significant, ΔR^2 values were examined to determine whether the significance was negligible ($\Delta R^2 < 0.035$), moderate ($0.035 \leq \Delta R^2 \leq 0.070$), or severe ($\Delta R^2 > 0.070$; Jodoin & Gierl, 2001).

MCFA is well suited to assess how items work together to influence scale means, including whether mean differences between groups are genuine or potentially a function of bias (Lavoie & Douglas, 2012; Milfont & Fischer, 2010). Given that the concern for gender bias emerged from observing mean differences between males and females on PSS scores, MCFA is an appropriate framework to help address the third goal (Lavoie & Douglas, 2012; Milfont & Fischer, 2010). This study is specifically interested in testing the loadings and intercepts for noninvariance across gender (i.e., scalar invariance [SI]; Sass, 2011). According

[1] Items were reverse scored for all analyses.

to Sass (2011), loadings and intercepts are used in MCFA to estimate latent factor means and when they fail to be invariant across groups latent mean differences may be obscured by bias, which impedes the attribution of group mean differences to genuine differences on the latent variable(s). Although the emphasis here is on latent means and concern for gender bias emerged from observed means, given the popularity of latent variable modeling, researchers are likely to still be interested in the findings. Furthermore, Lavoie and Douglas (2012) imply that findings regarding bias in latent means are generalizable to observed means as well.

Testing for SI proceeded in four steps as outlined by Sass (2011). First, the factor solution selected in the initial ordinal CFA was fit in males and females separately and examined for model-data fit to identify whether the factor structure is invariant across gender (i.e., configural invariance [CI]; Milfont & Fischer, 2010). Second, assuming the model fit the data well in males and females separately, the same factor model was simultaneously estimated across males and females with loadings and intercepts free to vary between groups. This CI model serves as a base for testing increasingly restrictive models (Sass, 2011). Third, loadings that were free in the CI model were constrained to be equal across males and females in the metric invariance (MI) model and tested for evidence of noninvariance (Sass, 2011). Testing the invariance of the loadings identifies whether the relationship between items and the factor(s) are equivalent across gender (Lavoie & Douglas, 2012; Milfont & Fischer, 2010). Fourth, assuming MI held, factor loadings and intercepts were constrained to be equal across groups in the SI model and tested against the MI model for noninvariance (Sass, 2011). Constraining the intercepts identifies whether males and females treat the Likert-type scale in a similar way (Milfont & Fischer, 2010). Note that the covariance between the latent factors, latent variances, and error terms were freely estimated in each model.

Metric and scalar models were tested for noninvariance using both χ^2 difference (χ^2_{diff}) testing and changes in CFIs between models (ΔCFI). The invariance testing outlined above produced a series of nested models whose χ^2 values and degrees of freedom were subtracted and tested for significant differences (i.e., noninvariance; Loehlin, 2004). That being said, χ^2_{diff} testing can be overly sensitive and may display problematic type I error rates in large samples and complex models, for example (Cheung & Rensvold, 2002). As a result, this study followed the recommendation of Cheung and Rensvold (2002), Sass (2011), and Chen (2007) that multiple criteria should be used to test for noninvariance. Cheung and Rensvold examined 20 goodness-of-fit (GFI) indices in a simulation study and found only ΔCFI and two additional fit indices were independent of sample size and model complexity and recommended a cutoff of $\Delta\text{CFI} \geq 0.01$ as evidence of noninvariance. Chen came to a similar conclusion and recommended the same criteria, while Sass found the $\Delta\text{CFI} < 0.01$ criteria to work well in ordinal MCFA. Thus, this same criterion was used in the present study.

Results

Ordinal CFA

Polychoric correlations for the PSS-10 are displayed separately for females and males in Table 1. The one factor solution did not fit the data well: Satorra-Bentler $\chi^2 = 898.945$ ($df = 35$, $p < .001$); CFI = 0.932; TLI = 0.913; RMSEA = 0.141 (confidence interval [CI]: 0.134 to 0.150). The orthogonal two-factor solution produced mixed results with respect to the fit criteria: Satorra-Bentler $\chi^2 = 372.138$, $df = 35$, $p < .001$; CFI = 0.974; TLI = 0.966; RMSEA = 0.0804 (CI: 0.0804 to 0.0966). Although the CFI and TLI suggest the model fit the data well, the RMSEA is a bit high. In contrast, the two-factor oblique solution fit the data well: $\chi^2 = 136.133$, $df = 34$, $p < .001$; CFI = 0.992; TLI = 0.989; RMSEA = 0.0493 (CI: 0.0408 to 0.0582). The standardized correlation coefficient between the latent factors was .76. As a result of these findings the unidimensional and orthogonal two-factor models were rejected for the oblique two-factor model. IRT and OLR DIF analyses were subsequently conducted on each subscale separately and the oblique two-factor model was specified in the SI testing.

IRT Analysis

As indicated previously, item-level fit was assessed visually by inspecting item-fit plots generated by MODFIT. Because this produced five separate graphs for each of the 10 items, only fit plots from one item ("Been Angered") are provided and displayed in Figure 1 as an example. Although, no other item showed as much deviation in the empirical probability curves from the pre-

Table 1
Interitem Polychoric Correlations of the PSS-10

Item	1	2	3	4	5	6	7	8	9	10
1. Been upset	1	.509	.447	.176	.249	.444	.184	.336	.585	.546
2. Unable to control	.597	1	.562	.411	.454	.564	.242	.508	.534	.690
3. Nervous and stressed	.521	.615	1	.246	.292	.471	.109	.367	.552	.612
4. Felt confident	.336	.443	.400	1	.660	.337	.425	.674	.286	.447
5. Going your way	.392	.548	.443	.657	1	.340	.381	.689	.289	.478
6. Could not cope	.394	.539	.509	.373	.433	1	.147	.450	.399	.653
7. Control irritations	.304	.348	.310	.478	.458	.307	1	.462	.114	.298
8. On top of things	.338	.560	.487	.576	.694	.508	.503	1	.358	.572
9. Been angered	.540	.538	.487	.311	.388	.457	.239	.383	1	.602
10. Couldn't overcome	.544	.636	.605	.421	.501	.616	.365	.531	.605	1

Note. PSS-10 = 10-item Perceived Stress Scale. Interitem polychoric correlations for males ($n = 540$) are presented above the diagonal, and interitem polychoric correlations for females ($n = 696$) are presented below the diagonal.

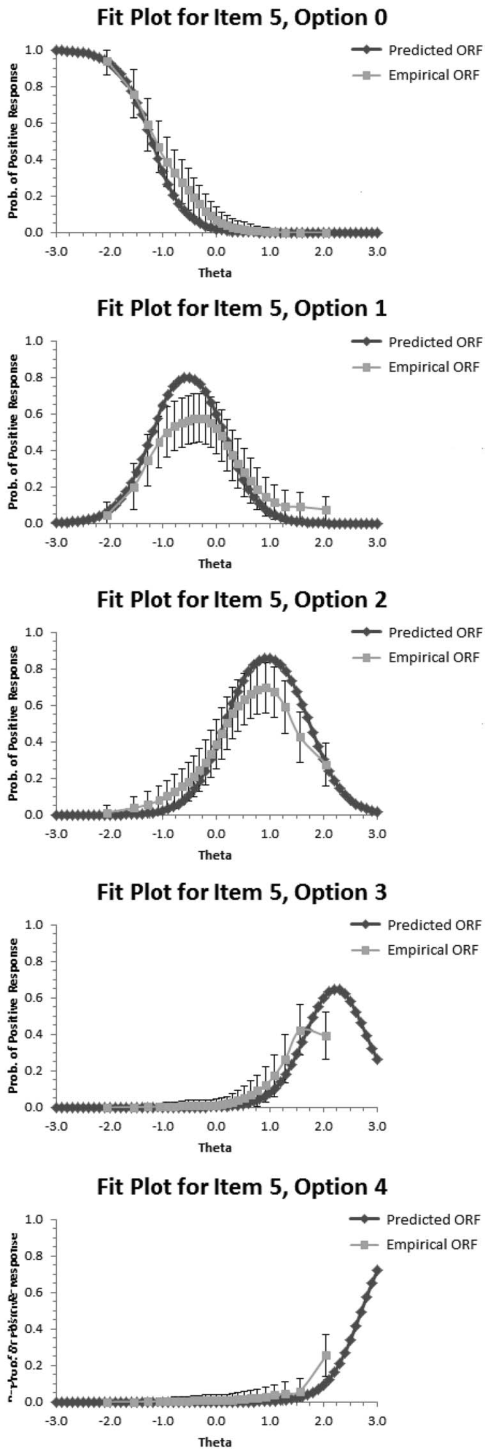

Figure 1. Empirical and predicted option response functions (ORF) with 95% error bars for Item 5 ("Been Angered"). Prob. = probability.

dicted probability curves than Item 5, inspection of Figure 1 shows that the predicted probability curves generally fall within the error bars of the empirical probability curves. The exception is category 1 (*Almost Never*) where we see some minor deviation of the predicted curve from the error bars of the observed curve for θ values approximately between −0.75 and 0.00. Still, the evidence suggests that Item 5 fits the GRM well. Overall, all items for both subscales fit the GRM well.

Figure 2 displays the trace lines for each item of the PHS and PSES subscales and Table 2 provides the estimated parameters for both subscales. On the PSES subscale, the item threshold parameters show little variation with δ_1 estimates ranging from −0.505 to –1.507, δ_2 estimates ranging from 0.376 to 0.897, δ_3 estimates ranging from 1.655 to 1.895, and δ_4 estimates ranging from 2.31 to 2.844. The α_j estimates suggest that all the items are highly discriminating (de Ayala, 2009; Estrada, Probst, Brown, & Graso, 2011). On the PHS subscale, the item threshold parameters are also fairly homogenous. The estimates for the lowest threshold parameters range from –1.587 to −0.183, δ_2 values range from 0.086 to 0.834, δ_3 values range from 1.628 to 2.117, and δ_4 estimates range from 2.582 to 3.256. As with the PSES, the α_j estimates suggest that all items are highly discriminating (de Ayala, 2009; Estrada et al., 2011). However, Item 6 ("Couldn't Overcome") has an exceptionally high discrimination estimate ($\alpha = 3.106$), which may indicate a problem with the item (Estrada et al., 2011).

Although the theoretical range for α_j is $-\infty$ to ∞, negative values and estimates larger than 3.00 are generally considered problematic (Baker & Kim, 2004; Estrada et al., 2011). In the latter case, α_j values larger than 3.00 are considered too good to be true and draw suspicion in applied contexts where values greater than 2.5 are unusual (Baker & Kim, 2004; Steinberg & Thissen, 1996). However, α_j estimates higher than 3.00 have been seen in the literature before (e.g., Estrada et al., 2011). Masters (1988) argued that an exceptionally high α_j estimate is more likely a product of a secondary (i.e., nuisance) factor in the form of item bias favoring respondents high on the latent continuum over that of respondents low on the continuum, giving the impression that an item is more discriminating between respondents of high and low latent levels than it actually is. Contrary to Masters (1988), however, Zhang (2008) found in a simulation study that α_j estimates tend to be an underestimate when influenced by a secondary factor. An investigation into Item 6 for DIF across high and low scorers (i.e., respondents above and below the mean) did not produce evidence of bias ($\Delta G^2 = 0.489$, $p = .484$). Trace lines for Item 6 displayed in Figure 2 are not diagnostic of any problems with the item (e.g., method effects; Steinberg & Thissen, 1996), and Item 6 fit the data well. Thus, the evidence in the present analysis suggests α_j may be accurate for Item 6.

Figure 4 displays the item and total test information curves for the PHS and PSES. Both the PHS and PSES exhibit a wide estimation range. On the PHS, the test information function suggests the subscale predominantly provides information for location estimates between –2.0 and 4.0. Respondents whose true locations are outside of these bounds are estimated with less reliability. The estimation ranges for the six PHS items are fairly homogenous with the exception of Item 6. As expected, the large contribution of Item 6 to the estimation of perceived helplessness is accompanied by a restricted estimation range relative to the other items (de Ayala, 2009). Most of the information Item 6 contributes appears to be over the interval –1.0 and 3.0. On the PSES, the test information function suggests the subscale predominantly provides reliable estimation of respondents' perceived self-efficacy over the interval –2.0 and 3.5. Examination of the item information curves

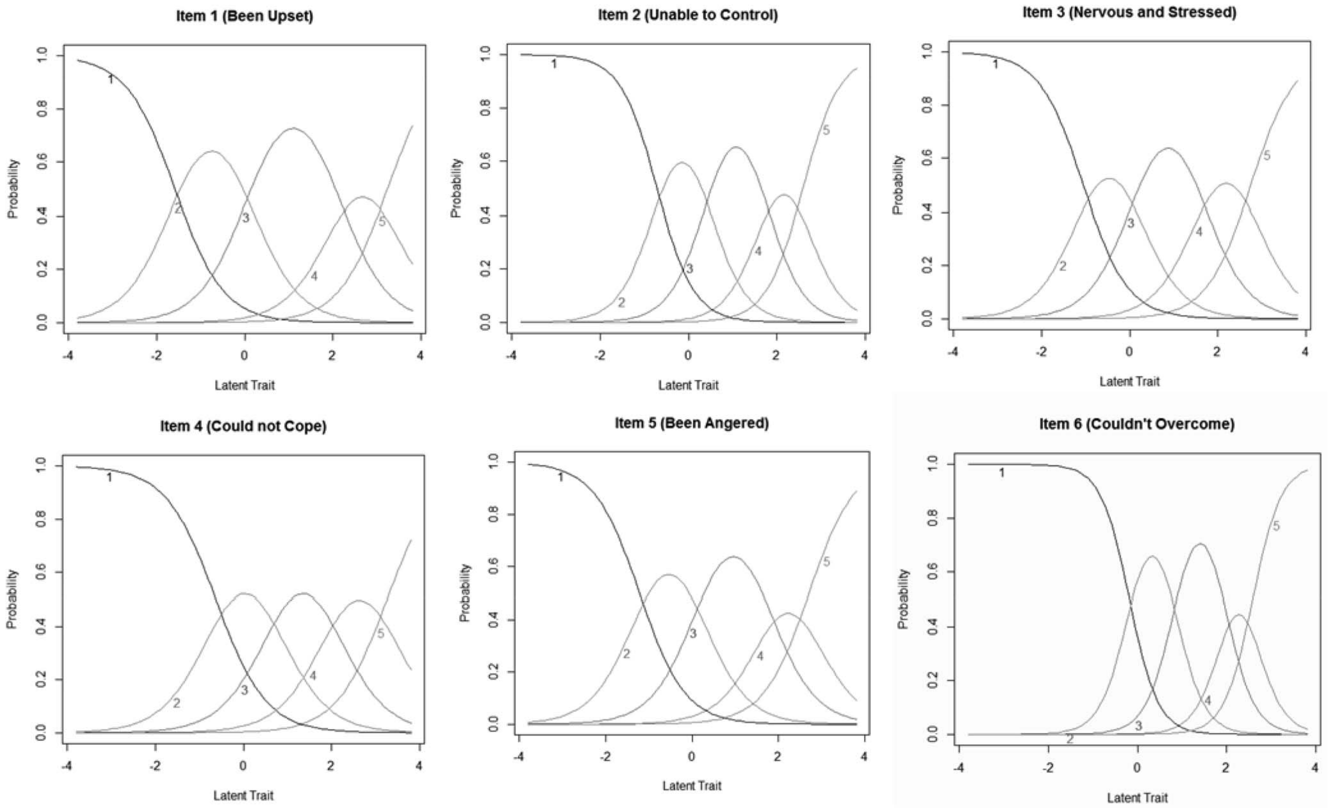

Figure 2. Individual trace lines of the six Perceived Helplessness Subscale (PHS) items.

reveals the estimation ranges of the four items are fairly homogenous.

Measurement Invariance

DIF. Consistent with previous studies (e.g., Lavoie & Douglas, 2012), on the PHS females tended to report higher scores (M = 13.73, SD = 4.36) than males (M = 12.75, SD = 4.16), while on the PSES females tended to report slightly lower scores (M = 8.87, SD = 2.88) than males (M = 9.00, SD = 3.00). As expected, independent samples t tests revealed a statistically significant difference between the means of males and females on the PHS (t = 4.001, df = 1234, $p < .001$) and not on the PSES (t = 0.747, df = 1234, p = .455). However, analyses (see Table 2) provided no evidence of uniform or nonuniform DIF for any of the items for

Table 2
Results of the IRT and DIF Analyses

	IRT parameters					Nonuniform DIF			Uniform DIF		
PSS-10 subscale and item	α	δ_1	δ_2	δ_3	δ_4	χ^2	p	ΔR^2	χ^2	p	ΔR^2
PHS											
1. Been upset	1.814	−1.587	0.086	2.117	3.236	0.16	.689	.000	0.02	.888	.000
2. Unable to control	2.4	−0.715	0.429	1.728	2.586	3.54	.060	.001	11.7	.000**	.001
3. Nervous and stressed	1.991	−1.055	0.115	1.628	2.745	0.56	.454	.000	4.6	.032*	.001
4. Could not cope	1.75	−0.628	0.69	2.008	3.256	0.10	.752	.000	13.5	.000**	.004
5. Been angered	1.89	−1.218	0.152	1.747	2.706	2.42	.120	.000	8.46	.003**	.003
6. Couldn't overcome	3.106	−0.183	0.834	1.964	2.582	0.00	.999	.000	0.08	.777	.000
PSES											
1. Felt confident	2.36	−0.505	0.897	1.833	2.31	0.42	.517	.013	40.24	.000**	.013
2. Going your way	2.953	−1.034	0.376	1.712	2.464	3.98	.460	.001	99.46	.000**	.032
3. Control irritations	1.335	−1.507	0.471	1.895	2.844	0.22	.639	.000	29.18	.000**	.000
4. On top of things	2.983	−0.987	0.55	1.655	2.544	0.84	.359	.000	55.7	.000**	.018

Note. IRT = item response theory; DIF = differential item functioning; PSS-10 = 10-item Perceived Stress Scale; PHS = Perceived Helplessness Subscale; PSES = Perceived Self-Efficacy Subscale.
* $p < .05$. ** $p < .01$.

either subscale. None of the items exhibited statistically significant nonuniform DIF, but as expected most items showed statistically significant uniform DIF due to the large sample size (Cheung & Rensvold, 2002). Items flagged as significant were inspected further for evidence of DIF using ΔR^2; however, none of the items in either the PHS or PSES exhibited moderate or severe uniform DIF. Figure 3 shows that only Item 2 of the PSES ("Going Your Way") came close to the 0.035 criteria for moderate DIF ($\Delta R^2 = 0.032$). The remaining items had much lower ΔR^2 levels.

MCFA. Testing for SI with ordinal indicators produced mixed results. The oblique two-factor solution selected in the initial ordinal CFA was fit in males and females separately. The model fit the data well in males: Satorra-Bentler $\chi^2 = 90.949$, $df = 34$, $p < .001$; CFI = 0.989; TLI = 0.986; RMSEA = 0.056 (CI: 0.042 to 0.070). The model also fit the data well in females: Satorra-Bentler $\chi^2 = 90.511$, $df = 34$, $p < .001$; CFI = 0.993; TLI = 0.990; RMSEA = 0.049 (CI: 0.037 to 0.061). Since these findings suggest CI holds, this study proceeded with the analyses by fitting the base model. As expected, the model fit the data well: Satorra-Bentler $\chi^2 = 181.477$, $df = 68$, $p < .001$; CFI = 0.992; TLI = 0.989; RMSEA = 0.052 (CI: 0.043 to 0.061). The MI model was considered next, which fit the data well: Satorra-Bentler $\chi^2 = 206.478$, $df = 76$, $p < .001$; CFI = 0.989; TLI = 0.990; RMSEA = 0.053 (CI: 0.044 to 0.062). The ΔCFI between the CI and MI models was < 0.01 but the Scaled χ^2_{diff} test was significant (Scaled $\chi^2_{\text{diff}} = 30.29$, $df = 8$, $p < .01$), which suggests the factor loadings may not be invariant across gender (Bryant, 2013). However, given the limitations of the χ^2_{diff} discussed above, the results of the ΔCFI, and the good fit of the MI model, the significant χ^2_{diff} does not justify rejection of the null hypothesis that the factor loadings are invariant. Thus, the analysis proceeded with the SI test assuming factor loadings are invariant. Again, the model fit the data well: Satorra-Bentler $\chi^2 = 266.107$, $df = 87$, $p < .001$; CFI = 0.987; TLI = 0.986; RMSEA = 0.058 (CI: 0.050 to 0.066). The ΔCFI between the MI and SI models was < 0.01 but the Scaled χ^2_{diff} test was significant again (Scaled $\chi^2_{\text{diff}} = 76.59$, $df = 11$, $p < .01$), which suggests the intercepts may not be invariant. Also again, given the limitations of the χ^2_{diff}, the results of the ΔCFI, and the good fit of the SI model, the significant χ^2_{diff} does not justify rejection of the null hypothesis.

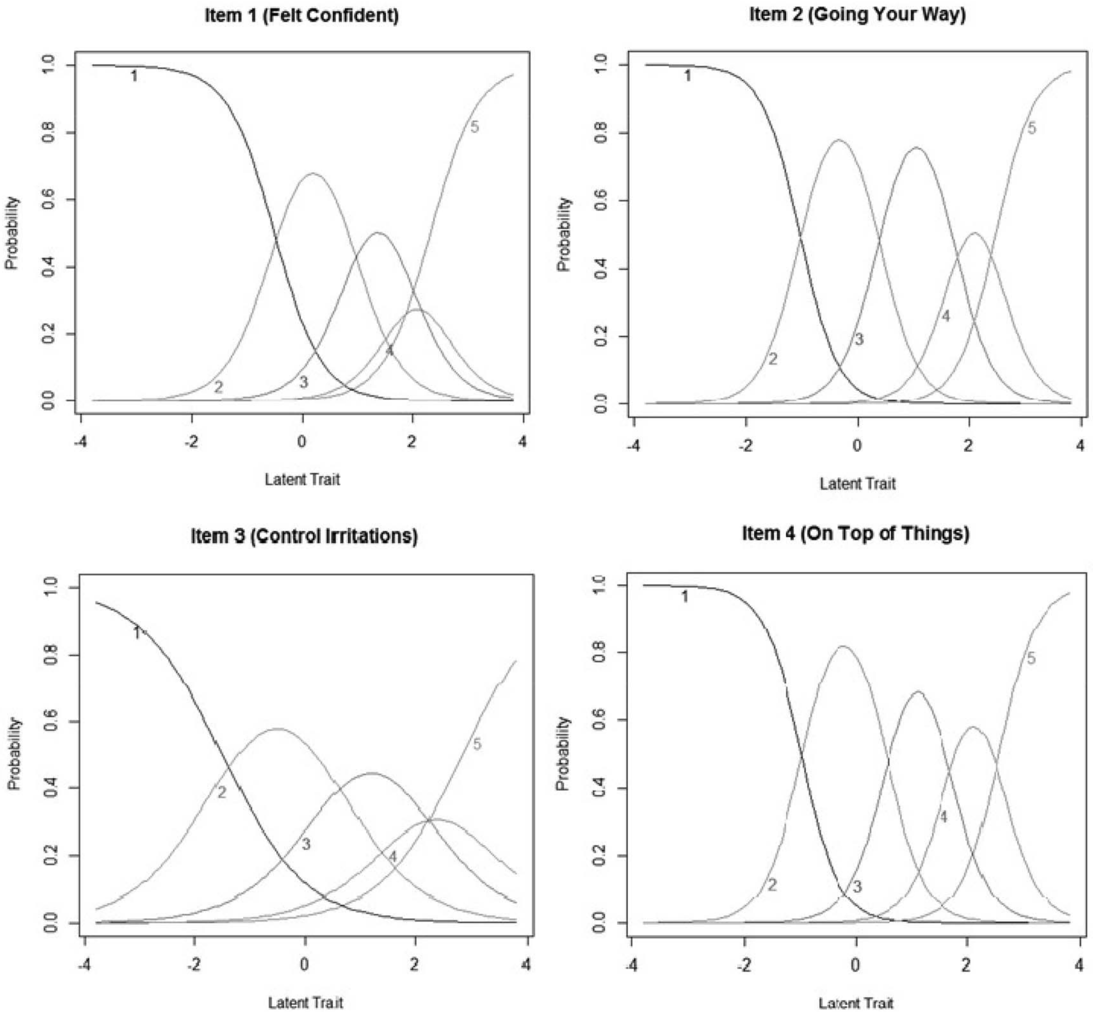

Figure 3. Individual trace lines of the four Perceived Self-Efficacy Subscale (PSES) items.

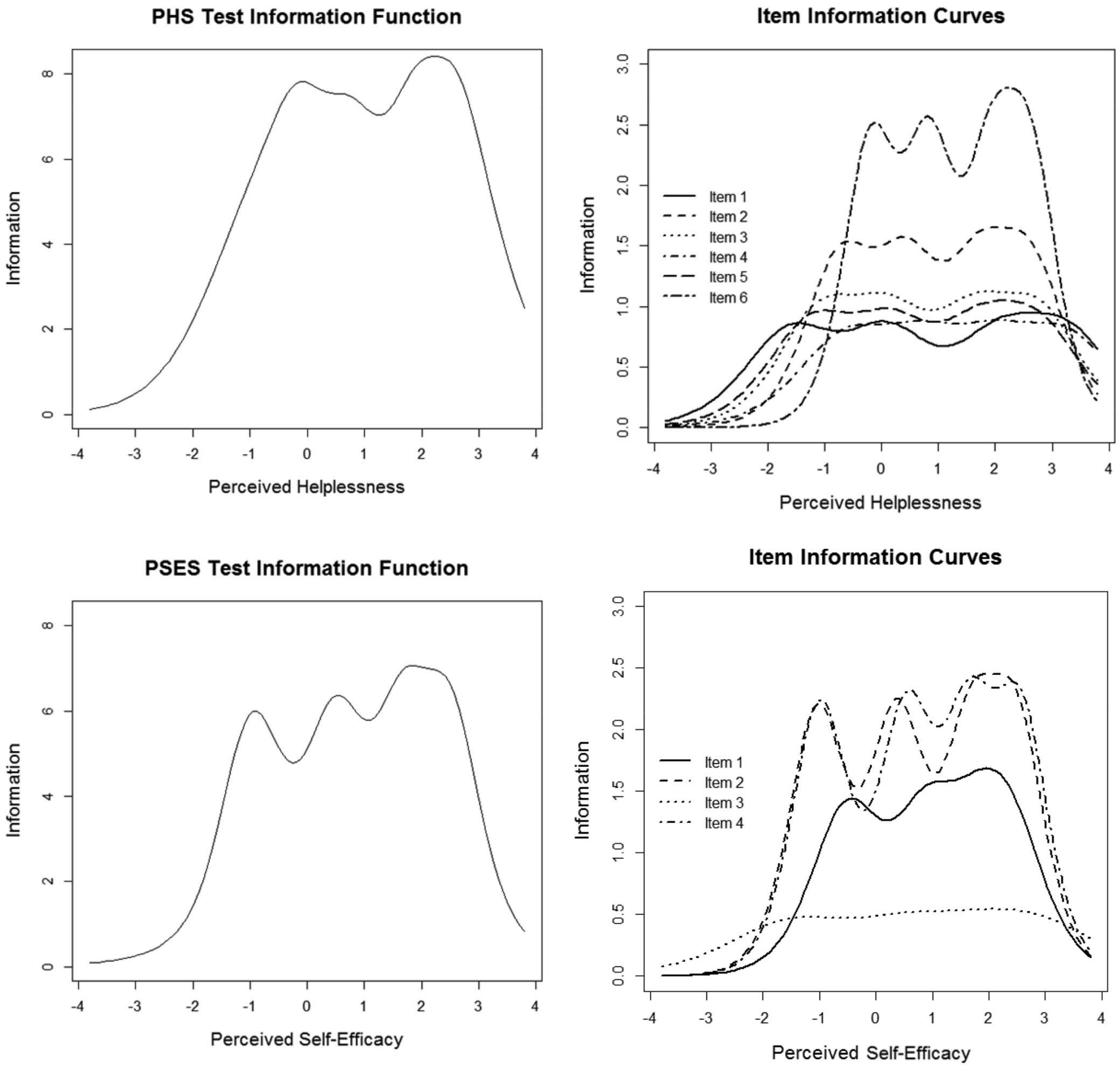

Figure 4. Total test and item information curves. PHS = Perceived Helplessness Subscale; PSES = Perceived Self-Efficacy Subscale.

According to Sass (2011), mixed results in invariance testing similar to those seen in this study suggests that the latent means may be subject to some bias but calls into question the practical significance of the violation. Thus, group comparisons of the latent factor means introduced in the SI model are likely to still be substantively meaningful. In the SI model the latent means of the female group were constrained to zero but freed in the male group. The latent factor mean of perceived helplessness was 0.143 units lower in males compared to females, while the latent factor mean of perceived self-efficacy was 0.003 units lower in males compared to females. Consistent with observed mean difference testing elsewhere in this study, the latent mean difference between males and females was significant on the perceived helplessness factor ($z = -12.573$, $p < .01$) and not on the perceived self-efficacy factor ($z = -0.095$, $p = .924$).

Discussion

The PSS has emerged as the most popular measure of perceived stress (Karam et al., 2012). Despite this popularity, a review of the literature reveals three significant gaps: (a) There is some debate as to whether a one or a two-factor model best describes the relationships among the PSS-10 items, (b) little information is available on the performance of the items on the scale, and (c) it is unclear whether PSS-10 scores are subject to gender bias. These gaps were addressed in this study using a sample of 1,236 adults from the MIDUS II study (Ryff et al., 2004–2009).

This study compared the original unidimensional factor structure hypothesized by Cohen and Williamson (1988), a correlated two-factor model, and an orthogonal two-factor model using ordinal CFA. In the latter two models the six negatively phrased PSS-10 items were loaded onto the perceived helplessness factor and the remaining four positively phrased items were loaded onto the perceived self-efficacy factor. The unidimensional and orthogonal models did not fit the data well, but the correlated two-factor model did. Thus, consistent with previous research, findings from the ordinal CFA suggests two factors—perceived helplessness and perceived self-efficacy—underlie the PSS-10 (e.g., Hewitt et al., 1992; Lavoie & Douglas, 2012). Additionally, Cohen and Williamson (1988) had asserted that the second factor is "irrelevant" in the measurement of perceived stress (p. 43). However, the critical role of the covariance between the two latent factors to the fit of the model suggests perceived self-efficacy is indispensable to the measurement of perceived stress including in nonclinical samples. As a result, the rest of the analyses focused on the two subscales (i.e., PHS and PSES) of the PSS-10 rather than the PSS-10 as a whole.

Results from an IRT analysis addressing item-level performance suggest both the PHS and PSES items generally follow the GRM well. Parameterization of the items indicate both the PHS and PSES effectively estimate a wide range of values along the latent continuum and discriminate well between respondents of differing latent levels. Nevertheless, researchers and clinicians who use the PSS-10 should be aware that reliable measurement of perceived stress becomes untenable as the degree of perceived helplessness becomes increasingly low and the degree of perceived self-efficacy becomes increasingly high. Prior to this study, the estimation capacity of each item at different points along the latent continuum had not been described, and it was unknown whether one or more of the items exhibit misfit.

Finally, measurement invariance tests suggest that PSS-10 scores may not be substantially affected by gender bias. Concern about gender bias emerged because female respondents tend to report higher levels of perceived stress than males on the PSS-10 (e.g., Cohen & Williamson, 1988; Lavoie & Douglas, 2012). Previous studies that addressed the issue found conflicting results. For example, Gitchel et al. (2011) found evidence of gender bias in four of the PSS-14's items, while Lavoie and Douglas (2012) found no evidence of gender bias in the PSS-14. The present study used ordinal logistic regression to test the items of each subscale for uniform and nonuniform DIF across gender while ordinal MCFA was used to test whether the PHS and PSES scale means may be subject to gender bias. DIF analyses provided no evidence for the view that the PHS and PSES items are subject to gender bias. Although the findings were mixed, the SI results suggested that the tendency of female respondents to report higher levels of perceived helplessness than males on the PSS-10 is substantively meaningful.

Limitations and Future Directions

The psychometric findings described in this study should be interpreted in light of a number of limitations. First, findings are generalizable only to populations the sample represents. In particular, participants were adults between the ages of 34 and 83 years, the sample did not reflect the full ethnic distribution in the United States, and participants had higher levels of education than the general U.S. population (Morozink et al., 2010). Thus, generalizability of the findings in this study to other respondents is unknown. Future research will need to validate the PSS-10's psychometric properties on diverse samples—especially in light of the concern that the PSS-10 may be biased by sample characteristics other than gender (e.g., race; Sharp et al., 2007). Second, although the general conclusion drawn in this study is that the PHS and PSES are not substantially affected by gender bias the issue may not be fully resolved. As seen between studies and even within this study, different methods of testing for invariance can produce different results and addressing these discrepancies in future studies will likely benefit both the study and application of the PSS-10 in research and clinical settings.

Implications

In spite of the limitations, the present study adds to the knowledge of the PSS-10's psychometric properties. The findings have at least three important implications. First, this study suggests inferences made based upon the PSS-10's scores are valid, as long as the instrument is used properly. That being said, second, researchers and clinicians must attend to the multidimensional nature of the PSS-10. Since multidimensionality can leave interpretations of test scores ambiguous, PSS-10 scores based upon all 10 items will be difficult to interpret and may not be valid (Ackerman, 1989). Finally, this study helps refine our understanding of stress. In particular, it supports a conceptualization of stress as involving both perceived helplessness and perceived self-efficacy as critical, interrelated elements. Interestingly, Lazarus's original stress model was revised years later and now emphasizes the role of perceived eustress alongside perceived distress (Golden-Kreutz, Browne, Frierson, & Andersen, 2004).

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13,* 113–127. http://dx.doi.org/10.1177/014662168901300201

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Barbosa-Leiker, C., Kostick, M., Lei, M., McPherson, S., Roper, V., Hoekstra, T., & Wright, B. (2013). Measurement invariance of the Perceived Stress Scale and latent mean differences across gender and time. *Stress and Health, 29,* 253–260.

Brim, O. G., Ryff, C. D., & Kessler, R. C. (2004). *How healthy are we? A national study of well-being at midlife*. Chicago, IL: Chicago Press.

Bryant, F. B. (2013). [EXCEL macro file for conducting Bryant-Satorra scaled difference chi-square test via LISREL 8]. Available from the author.

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. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Christensen, R. H. B. (2011). *Ordinal—regression models for ordinal data* (R package Version 2010.12–15). Retrieved from http://www.cran.r-project.org/package=ordinal/

Cohen, S. (2013). *PSS: Frequently asked questions*. Retrieved from http://www.psy.cmu.edu/~scohen/

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. M. (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.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Estrada, A. X., Probst, T. M., Brown, J., & Graso, M. (2011). Evaluating the psychometric and measurement characteristics of a measure of sexual orientation harassment. *Military Psychology, 23,* 220–236. http://dx.doi.org/10.1080/08995605.2011.559394

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.

Gitchel, W. D., Roessler, R. T., & Turner, R. C. (2011). Gender effect according to item directionality on the Perceived Stress Scale. *Rehabilitation Counseling Bulletin, 55,* 20–28. http://dx.doi.org/10.1177/0034355211404567

Golden-Kreutz, D. M., Browne, M. W., Frierson, G. M., & Andersen, B. L. (2004). Assessing stress in cancer patients: A second-order factor analysis model for the Perceived Stress Scale. *Assessment, 11,* 216–223. http://dx.doi.org/10.1177/1073191104267398

Hewitt, P. L., Flett, G. L., & Mosher, S. W. (1992). The Perceived Stress Scale: Factor structure and relation to depression symptoms in a psychiatric sample. *Journal of Psychopathology and Behavioral Assessment, 14,* 247–257. http://dx.doi.org/10.1007/BF00962631

Holgado-Tello, F. P., Carrasco-Ortiz, M. A., Gándara, M., & Moscoso, S. C. (2009). Factor analysis of the Big Five Questionnaire using polychoric correlations in children. *Quality & Quantity: International Journal of Methodology, 43,* 75–85.

Holgado-Tello, F. P., Chacón-Mascoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity: International Journal of Methodology, 44,* 153–166.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329–349. http://dx.doi.org/10.1207/S15324818AME1404_2

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.

Karam, F., Berard, A., Sheehy, O., Huneau, M., Briggs, G., Chambers, C., . . . OTIS Research Committee. (2012). Reliability and validity of the 4-item Perceived Stress Scale among pregnant women: Results from the OTIS antidepressants study. *Research in Nursing & Health, 35,* 363–375. http://dx.doi.org/10.1002/nur.21482

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36,* 399–419. http://dx.doi.org/10.1177/0146621612446170

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

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation, 12,* 1–11.

Leung, D. Y., Lam, T., & Chan, S. S. (2010). Three version of the Perceived Stress Scale: Validation in a sample of Chinese cardiac patients who smoke. *BMC Public Health, 10,* 513–520. http://dx.doi.org/10.1186/1471-2458-10-513

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Erlbaum.

Love, G. D., Seeman, T. E., Weinstein, M., & Ryff, C. D. (2010). Bioindicators in the MIDUS national study: Protocol, measures, sample, and comparative context. *Journal of Aging and Health, 22,* 1059–1080. http://dx.doi.org/10.1177/0898264310374355

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25,* 15–29. http://dx.doi.org/10.1111/j.1745-3984.1988.tb00288.x

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3,* 111–121.

Mitchell, A. M., Crane, P. A., & Kim, Y. (2008). Perceived stress in survivors of suicide: Psychometric properties of the Perceived Stress Scale. *Research in Nursing & Health, 31,* 576–585. http://dx.doi.org/10.1002/nur.20284

Morozink, J. A., Friedman, E. M., Coe, C. L., & Ryff, C. D. (2010). Socioeconomic and psychosocial predictors of interleukin-6 in the MIDUS national sample. *Health Psychology, 29,* 626–635. http://dx.doi.org/10.1037/a0021360

Ramírez, M., & Hernanzez, R. (2007). Factor structure of the Perceived Stress Scale (PSS) in a sample from Mexico. *The Spanish Journal of Psychology, 10,* 199–206. http://dx.doi.org/10.1017/S1138741600006466

Reis, R. S., Hino, A. A. F., & Rodriguez-Anez, C. R. 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

Remor, E. (2006). Psychometric properties of a European Spanish version of the Perceived Stress Scale (PSS). *The Spanish Journal of Psychology, 9,* 86–93. http://dx.doi.org/10.1017/S1138741600006004

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software, 17,* 1–25.

Roberti, J. W., Harrington, L. N., & Storch, E. A. (2006). Further psychometric support for the 10-item version of the Perceived Stress Scale. *Journal of College Counseling, 9,* 135–147. http://dx.doi.org/10.1002/j.2161-1882.2006.tb00100.x

Ryff, C. D., Seeman, T., & Weinstein, M. (2004–2009). *National Survey of Midlife Development in the United States (MIDUS II): Biomarker project* (ICPSR29282-v3). Ann Arbor, MI: Inter-university Consortium for Political and Social Research. http://dx.doi.org/10.3886/ICPSR29282.v3

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34,* 100–114.

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

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., Graeff, A. D., Groenvold, M., . . . Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62,* 288–295. http://dx.doi.org/10.1016/j.jclinepi.2008.06.003

Sharp, L. K., Kimmel, L. G., Kee, R., Saltoun, C., & Chang, C. (2007). Assessing the Perceived Stress Scale for African American adults with asthma and low literacy. *Journal of Asthma, 44,* 311–316. http://dx.doi.org/10.1080/02770900701344165

Shewchuk, R. M., Elliott, T. R., MacNair-Semands, R. R., & Harkins, S. (1999). Trait influences on stress appraisal and coping: An evaluation of alternative frameworks. *Journal of Applied Social Psychology, 29,* 685–704. http://dx.doi.org/10.1111/j.1559-1816.1999.tb02019.x

Stark, S. (2002). MODFIT: Plot theoretical item response functions and examine the fit of dichotomous or polytomous IRT models to response data [computer program]. Department of Psychology, University of Illinois at Urbana–Champaign.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91,* 1292–1306. http://dx.doi.org/10.1037/0021-9010.91.6.1292

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in measurement of psychopathology. *Psychological Methods, 1,* 81–97. http://dx.doi.org/10.1037/1082-989X.1.1.81

Wang, W., & Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65,* 376–404. http://dx.doi.org/10.1177/0013164404268673

Zhang, B. (2008). Application of unidimensional item response models to test with items sensitive to secondary dimensions. *Journal of Experimental Education, 77,* 147–166. http://dx.doi.org/10.3200/JEXE.77.2.147-166

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Received March 23, 2013
Revision received June 23, 2014
Accepted June 26, 2014 ■