ELSEVIER

Behaviour Research and Therapy 41 (2003) 1411–1426

BEHAVIOUR RESEARCH AND THERAPY

www.elsevier.com/locate/brat

Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects?

Timothy A. Brown *

Boston University, Center for Anxiety and Related Disorders, 648 Beacon Street 6th Floor, Boston, MA 02215-2013, USA

Received 22 January 2003; accepted 2 February 2003

Abstract

The latent structure of the Penn State Worry Questionnaire (PSWQ) was evaluated with confirmatory factor analyses (CFAs) in 1200 outpatients with *DSM-IV* anxiety and mood disorders. Of particular interest was the comparative fit and interpretability of a two-factor solution (cf. Behaviour Research and Therapy 40 (2002) 313) vs. a one-factor model that specified method effects arising from five reverse-worded items. Consistent with prediction, the superiority of the one-factor model was demonstrated in split-sample CFA replications (*n*s = 600). Multiple-group CFAs indicated that the measurement properties of the PSWQ were invariant in male and female patients. In addition to their direct relevance to the psychometrics of the PSWQ, the results are discussed in regard to methodological considerations for using factor analytic methods in the evaluation of psychological tests.

Keywords: Confirmatory factor analysis; Psychometrics; Pathological worry; Generalized anxiety disorder; Method effects; Assessment of worry

The diagnostic definition of generalized anxiety disorder (GAD) was modified substantially in the revised third edition of the Diagnostic and Statistical Manual of Mental Disorders (*DSM-III-R*; American Psychiatric Association, 1987). In addition to its elevation from a residual category to a full disorder, GAD was defined in *DSM-III-R* by the key feature of excessive worry. The fundamental nature of this feature was further underscored in the fourth edition of *DSM* (*DSM-IV*, American Psychiatric Association, 1994) which added the requirement that the worry be perceived as uncontrollable (Brown, Barlow, & Liebowitz, 1994). Changes in the formal nosology

* Tel.: +1-617-353-9610; fax: +1-617-353-9609.
E-mail address: tabrown@bu.edu (T.A. Brown).

0005-7967/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0005-7967(03)00059-7

led to a burgeoning empirical interest in pathological worry and the need for psychometrically sound measures of this construct. Since its publication in 1990, the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990; Molina & Borkovec, 1994) has emerged as the most widely used self-report measure of worry and GAD. Indeed, over the past 12 years, the PSWQ has been used in most psychosocial studies of worry and GAD including treatment outcome trials (e.g. Barlow, Rapee, & Brown, 1992; Borkovec & Costello, 1993; Ladouceur et al., 2000), analog worry research (e.g. Coles, Mennin, & Heimberg, 2001; Wells & Papageorgiou, 1998), laboratory investigations (e.g. Butler, Wells, & Dewick, 1995), GAD psychopathology studies (e.g. Beck, Stanley, & Zebb, 1996; Brown, Marten, & Barlow, 1995), and nosological evaluations of the discriminant validity of GAD in relation to overlapping anxiety and mood disorders (e.g. Brown, Chorpita, & Barlow, 1998; Brown, Moras, Zinbarg, & Barlow, 1993; Starcevic, 1995). Despite the popularity and empirical utility of the PSWQ, researchers who have examined its latent structure have reported inconsistent findings.

The PSWQ consists of 16 items to which participants respond using a scale from 1 (not at all typical of me) to 5 (very typical of me). Eleven items are worded in the direction of pathological worry (e.g. 'Once I start worrying, I cannot stop'), and the remaining five items are 'stated in a reversed fashion to reduce the effects of acquiescence' (p. 488, Meyer et al., 1990; e.g. 'I never worry about anything'). After reverse-scoring these five items, a single total score is created by summing the 16 items; thus, higher PSWQ scores reflect greater levels of pathological worry. In the initial psychometric study by Meyer, Miller, Metzger and Borkovec (1990) that relied on a predominantly nonpatient sample, the PSWQ was shown to have high internal consistency (e.g. $\alpha = 0.93$), temporal stability (rs ≥ 0.74 over intervals of 2 to 10 weeks), and favorable convergent and discriminant validity. These results were upheld in a clinical evaluation ($n = 436$) by Brown, Antony and Barlow (1992) who also reported high levels of internal consistency, concurrent validity, and the ability of the PSWQ to differentiate patients with GAD from patients with other anxiety and mood disorders. In both the Meyer et al. (1990) and Brown et al. (1992) studies, principal components analysis extracted two factors with unreduced eigenvalues greater than 1.0 (e.g. 8.17 and 1.23, accounting for 51.1 and 7.7% of the variance, respectively; Brown et al., 1992). Nevertheless, a unidimensional solution was retained in both investigations on the basis of scree tests and substantive issues (see below).

Researchers who have conducted exploratory latent structural analyses of the PSWQ subsequent to Meyer et al. (1990) and Brown et al. (1992) have consistently reported two-factor solutions (Beck, Stanley, & Zebb, 1995; Stöber, 1995; van Rijsoort, Emmelkamp, & Vervaeke, 1999). In every case, the first factor was comprised of the 11 items phrased in the symptomatic direction, whereas the second factor consisted of the five reverse-worded items. Although having no apparent basis in prior theory, the second factor was often interpreted in these studies as having substantive meaning; that is, it represented a dimension of 'Absence of Worry'.

However, these results depict a common outcome in the psychometric evaluation of psychological tests; namely that when scales are developed to entail a mixture of positively and negatively worded items, factor analyses will frequently produce distinct factors whereby positively worded items load on one factor and reverse-worded items on another (e.g. Bagozzi, 1993; Bagozzi & Heatherton, 1994; Baumgartner & Steenkamp, 2001; Byrne, Fisher, Lamberth, & Mitchell, 1974; Carmines & Zeller, 1979; Marsh, 1992, 1996; Spector, Van Katwyk, Brannick, & Chen, 1997). In these scenarios, the key question arises as to whether the factors have substantive meaning or whether they are an artifact of response styles associated with the wording of the items (i.e. responses biases such as acquiescence; cf. Marsh, 1996). Specifically, does the observed multifactorial structure indicate the presence of more than one clinically and theoretically salient dimensions or is it the result of substantively irrelevant method effects? Resolution of this issue is critical to the process of scale development as it will have far-reaching implications to the future measurement and conceptualization of the construct.

Although exploratory factor analysis (EFA) often produces distinct factors comprised of positively and negatively worded components (e.g. Beck, Stanley & Zebb, 1995; Carmines & Zeller, 1979), this approach has little utility to determining the nature of these outcomes (e.g. an a priori structure cannot be imposed beyond specification of the number of latent factors, m; the m^2 restrictions used in EFA model identification preclude an analysis of error covariances). In contrast, confirmatory factor analysis (CFA) provides a very powerful and flexible framework for addressing such issues (Jöreskog, 1969, 1971). Unlike EFA, restrictions can be placed on the various parameter estimates (i.e. factor loadings, variances, covariances, residual variances) thereby resulting in a more parsimonious model. Thus, the estimation of latent factor measurement models is not limited to the assumption of uncorrelated uniqueness as in classical test theory (Lord & Novick, 1968) and the traditional common factor model (Thurstone, 1947). Rather, provided that the CFA model is identified (e.g. the number of elements in the input variance/covariance matrix exceeds the number of freely estimated parameters), correlated errors can be specified among the indicator residuals to reflect various types of method effects such as high content overlap, similar phrasings, differential susceptibility to demand characteristics, and carelessness or difficulty reading reverse-worded items (cf. Byrne, Shavelson, & Muthén, 1989; Cordery & Sevastos, 1993; Floyd & Widaman, 1995; Gerbing & Anderson, 1984; Schmitt & Stults, 1985; Spector, Van Katwyk, Brannick & Chen, 1997). In the case of congeneric models (i.e. multiple indicator measurement models in which indicators do not load on more than one latent factor; Jöreskog, 1971), these specifications are made under the assumption that the observed covariances among items loading on a common factor can not be accounted entirely by the underlying ('true score') dimension (i.e. some of this covariation is due to sources other than latent construct). Thus, in measurement models where method effects are present, it is important that the solution be fitted accordingly given the impact these error covariances will have on the resulting estimates of factor loadings, scale reliabilities, factor scores, and factor determinacies (Green & Hershberger, 2000; Raykov, 2001).

This discussion is nicely illustrated by the extensive psychometric literature on Rosenberg's (1965) Self-Esteem Scale (SES; e.g. Bachman & O'Malley, 1986; Carmines & Zeller, 1979; Marsh, 1986, 1996; Tomás & Oliver, 1999; Wang, Siegal, Falck, & Carlson, 2001). Early factor analytic research routinely produced two SES factors comprised of negatively and positively worded items (e.g. Bachman & O'Malley, 1986; Carmines & Zeller, 1979). Although the potential influence of method effects was considered in some of these studies (e.g. Bachman & O'Malley, 1986; Carmines & Zeller, 1979), other researchers interpreted the two factors as substantively meaningful (e.g. 'general positive self-evaluation' vs. 'transient negative self-evaluation'). In an attempt to reconcile the debate in this literature, Marsh (1996) presented a CFA approach to the structural evaluation of the SES, noting the many advantages of this strategy (e.g. ability to statistically compare competing factor structures and incorporate an error theory). Using data from a large, nationally representative sample, Marsh (1996) evaluated various measurement models corresponding to previously reported solutions (e.g. one-factor model without error covariances, two-factor models; cf. Carmines & Zeller, 1979) and correlated uniqueness (residual) models popularized in research on CFA approaches to multitrait-multimethod data (e.g. Marsh & Grayson, 1995; Tomás, Hontangas, & Oliver, 2000). Results indicated the superiority of a unidimensional solution ('global self-esteem') with method effects (correlated residuals) associated with the negatively worded items. Subsequent CFA studies have upheld Marsh's (1996) finding that the SES is defined by a single factor of global self-esteem with an error theory accounting for the covariances of the residuals of similarly worded items (Tomás & Oliver, 1999; Wang, Siegal, Falck & Carlson, 2001).

In addition to other limitations (use of an extraction method not based on the common factor model; cf. Fabrigar, Wegener, MacCallum, & Strahan, 1999; Widaman, 1993), studies that conducted exploratory analyses of the PSWQ used an analytic framework (principal components analysis) that precluded evaluation of competing explanations of the measure's latent structure. In the only published CFA-based evaluation to date, Fresco, Heimberg, Mennin and Turk (2002) examined the latent structure of the PSWQ in a large college student sample ($n = 788$). However, CFA was used only as a method to statistically compare the fit of one- and two-factor solutions. Indeed, based on the more recent results of Beck, Stanley and Zebb (1995); Stöber (1995), and van Rijsoort et al. (1999), the authors posited 'that an argument could be made for the appropriateness of the two-factor solution' (p. 315, Fresco et al., 2002). Consistent with this contention, a two-factor model, comprised of positively and negatively worded items, was found to provide a superior fit to the data. In addition to the factor defined by the 11 items phrased in the symptomatic direction (labeled 'Worry Engagement'), Fresco et al. (2002) ascribed substantive meaning to the factor consisting of the five reverse-worded items (labeled 'Absence of Worry'). Subsequent analyses were undertaken to examine the internal consistency, norms, and concurrent validity of the worry engagement and absence of worry factors.

However, the methods and results of Fresco et al. (2002) could be questioned because the competing models were specified without considering the vast literature (and empirical basis) on the role of method effects in measures with positively and negatively worded items (cf. Marsh, 1996). Moreover, the decision to accept the two-factor model over the unidimensional model was based solely on goodness of fit, without due consideration of the interpretability of this solution. Although the two-factor model provided a better fit to the data, the acceptability of this solution would also necessitate a compelling justification for the clinical, conceptual, or empirical meaningfulness of an 'absence of worry' dimension.

Thus, a key aim of the present study was to re-evaluate the latent structure of the PSWQ, guided by the prediction that a one-factor solution with an appropriate error theory would be empirically and conceptually more suitable than the two-factor solution reported by Fresco et al. (2002). In addition to shedding further light on the underlying structure of the PSWQ, some of the other advantages and methodological considerations in the use of CFA in psychological test development are addressed (e.g. evaluation of complex error structures, factor determinacy, scale reliability, form equivalence, measurement invariance, and population heterogeneity).

1. Method

1.1. Participants

The sample consisted of 1200 consecutive outpatient admissions to the Center for Anxiety and Related Disorders. Patients completed the PSWQ as part of their initial intake evaluation at the Center (consisting of a structured interview and a questionnaire battery). Women constituted the larger portion of the sample (61.5%); average age was 33.02 (S.D. = 11.02, range, 18–73). The sample was predominantly Caucasian (89.8%; African-American, 3.4%, Asian, 3.3%, Latino/Hispanic, 2.8%). Diagnoses were established with the Anxiety Disorders Interview Schedule for *DSM-IV*: Lifetime version (ADIS-IV-L; Di Nardo, Brown, & Barlow, 1994), a semi-structured interview designed to provide reliable diagnosis of the *DSM-IV* anxiety, mood, somatoform, and substance use disorders, and to screen for other conditions (e.g. psychotic disorders). A reliability study of a subset of this sample ($n = 362$) who had two independent administrations of the ADIS-IV-L indicated good-to-excellent interrater agreement for current disorders (range of κs = 0.67 to 0.86) (Brown, Di Nardo, Lehman, & Campbell, 2001). The sample breakdown of current clinical disorders (collapsing across principal and additional diagnoses) was: panic disorder with or without agoraphobia ($n = 500$), generalized anxiety disorder ($n = 266$), social phobia ($n = 487$), specific phobia ($n = 220$), obsessive-compulsive disorder ($n = 143$), major depression ($n = 336$), dysthymic disorder ($n = 95$), other anxiety/mood disorder (e.g. posttraumatic stress disorder, anxiety or depressive disorder NOS; $n = 192$).

1.2. Measure

The PSWQ is a 16-item questionnaire designed to assess excessive and uncontrollable worry (Meyer et al., 1990; Molina & Borkovec, 1994). Participants respond to items using a 1–5 scale, where 1 is ‘not at all typical of me’ and 5 is ‘very typical of me.’ After reverse-scoring five items, a total score is formed by summation (i.e. range of scores 16 to 80 with higher scores reflecting higher levels of worry).

1.3. Procedure and data analysis

Initial CFAs were conducted in two random subsamples (split sample 1 and 2, ns = 600) as a cross-validation strategy. The subsamples were then combined to conduct a full-sample CFA and to evaluate whether the PSWQ’s measurement properties were invariant in male and females. The variance–covariance matrices were analyzed using latent variable software programs and maximum-likelihood minimization functions (LISREL 8.51, Jöreskog & Sörbom, 2001; Mplus 2.02, Muthén & Muthén, 2001). Goodness of fit was evaluated using the root mean square error of approximation (RMSEA) and its 90% confidence interval (90% CI; cf. MacCallum, Browne, & Sugawara, 1996), standardized root mean square residual (SRMR), comparative fit index (CFI), and the Tucker–Lewis index (TLI). Acceptable model fit was defined by the following criteria: RMSEA (<0.08, 90% CI <0.08), SRMR (<0.05), CFI (>0.90), and TLI (>0.90). Multiple indices were used because they provide different information about model fit (i.e. absolute fit, fit adjusting for model parsimony, fit relative to a null model); used together, these indices provide a more conservative and reliable evaluation of the solution (cf. Jaccard & Wan, 1996). Most of the fitted models were nested; in these instances, comparative fit was evaluated by χ^2 difference tests (χ^2_{diff}) and by the interpretability of the solutions.

2. Results

2.1. Confirmatory factor analysis

Of particular interest was the comparative evaluation of the two-factor solution forwarded by Fresco et al. (2002) and a unifactorial model incorporating an error theory to reflect the method effect from the five reverse-worded items. However, one-factor solutions (with no residual covariances among the 16 items) were fit to the data first to serve as baseline models and to identify salient sources of misfit before proceeding to the competing models of interest. For instance, research has shown that even for measures containing items worded in the same direction, it is frequently necessary to specify correlated residuals that represent nonrandom measurement error due to other types of method effects (e.g. items with highly similar wordings or content; cf. Byrne, Shavelson & Muthén, 1989; Floyd & Widaman, 1995; Gerbing & Anderson, 1984).

Thus, a one-factor solution was fit to the split sample 1 ($n = 600$) data. As shown in Table 1, this model did not fit the data well (e.g. RMSEA = 0.09). Evaluation of localized areas of strain in this solution indicated that, in addition to not reproducing all of the covariance among reverse-worded items, there was strong evidence of a correlated residual between items 9 and 16 (modification index = 109.31, standardized expected parameter change = 0.24). Consideration of this outcome suggested that the covariance of these items that was unaccounted for by the latent factor was likely due to a method effect stemming from content overlap (‘As soon as I finish one task, I start to worry about everything else I have to do,’ ‘I worry about projects until they are done’). Accordingly, the one-factor model was refit to the data freely estimating the error covariance between items 9 and 16. This respecification resulted in a significant improvement in model fit, $\chi^2_{\text{diff}}(1) = 115.75$, $P < 0.001$ ($\theta_{16,9} = 0.24$, $P < 0.001$), although overall fit was still unsatisfactory (e.g. RMSEA = 0.081). Fit diagnostics revealed the existence of another significant error covariance (modification index = 50.56, standardized expected parameter change = 0.09) between two other highly similar items (items 7 and 15: ‘I am always worrying about something,’ ‘I worry all the time’). Relaxing this covariance also significantly improved the fit of the model, $\chi^2_{\text{diff}}(1) = 48.15$, $P < 0.001$ ($\theta_{17,5} = 0.09$, $P < 0.001$). Therefore, the correlated errors between items 9 and 16 and items 7 and 15 were freely estimated in subsequent models.

Next, the two-factor model that was obtained and accepted by Fresco et al. (2002) was fit to the data (‘Worry Engagement,’ ‘Absence of Worry’). Not surprisingly, the fit of this solution was superior to that of the preceding model, $\chi^2_{\text{diff}}(1) = 77.91$, $P < 0.001$, due to its ability to better reproduce the intercorrelations among the reverse-worded items. Indeed, all fit indices were consistent with good model fit (see Table 1), and both factors were defined by items with salient and statistically significant factor loadings (range of factor loadings were 0.64 to 0.85 and 0.43 to 0.73 for Worry Engagement and Absence of Worry, respectively). However, there were at least two compelling reasons to question the acceptability of this model. In addition to the aforementioned considerations regarding the tenuous substantive meaning of an ‘Absence of Worry’ factor,

Table 1
Confirmatory factor analyses of the Penn State Worry Questionnaire: Overall model fit

	χ^2	df	χ^2_{diff}	Δdf	RMSEA	(90% CI)	SRMR	CFI	TLI
Split sample 1 (n = 600)									
One factor	623.97	104			0.091	(0.084–0.098)	0.045	0.911	0.897
One factor with $\theta_{16,9}$ free[a]	508.22	103	115.75***	1	0.081	(0.074–0.088)	0.041	0.931	0.919
One factor with $\theta_{16,9}$, $\theta_{15,7}$ free[b]	460.07	102	48.15***	1	0.076	(0.069–0.084)	0.040	0.939	0.928
Two factors[c]	382.16	101	77.91***	1	0.068	(0.061–0.075)	0.034	0.952	0.943
One factor with method effects[d]	343.07	92	39.09***	9	0.067	(0.060–0.075)	0.031	0.957	0.944
Split sample 2 (n = 600)									
One factor	573.26	104			0.087	(0.080–0.094)	0.050	0.908	0.893
One factor with $\theta_{16,9}$ free[a]	510.06	103	63.20***	1	0.081	(0.074–0.088)	0.047	0.920	0.907
One factor with $\theta_{16,9}$, $\theta_{15,7}$ free[b]	464.26	102	45.80***	1	0.077	(0.070–0.084)	0.046	0.929	0.916
Two factors[c]	334.88	101	129.38***	1	0.062	(0.055–0.070)	0.035	0.954	0.945
One factor with method effects[d]	314.01	92	20.87**	9	0.063	(0.056–0.071)	0.032	0.956	0.943

χ^2_{diff}, nested χ^2 difference; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; SRMR, standardized root mean square residual; CFI, comparative fit index; TLI, Tucker-Lewis index. $^{**}P < 0.025$, $^{***}P < 0.001$.

[a] Correlated residuals between items 9 and 16.

[b] Correlated residuals between items 7 and 15.

[c] $\theta_{16,9}$ and $\theta_{15,7}$ freely estimated in the two-factor model and the hypothesized one-factor model.

[d] Covariances among the errors of reverse-worded items (items 1, 3, 8, 10, 11).

the two factors were highly correlated ($\phi_{2,1}$ = 0.87) suggesting poor discriminant validity of these dimensions and the possibility that a more parsimonious solution could be attained.

The underlying assumption of the hypothesized one-factor model is that most of the covariance among the 16 PSWQ items is due to the influence of a single latent dimension of worry (the common factor), but a more complex error theory is needed to account for the residual covariations (systematic error) among responses to reverse-worded items (i.e. the differential covariation of the five reverse-scored items is due to method effects rather than the existence of another meaningful construct). As shown in Table 1, all fit indices pointed to a good-fitting solution, and fit was superior to the two-factor model, $\chi^2_{\text{diff}}(9) = 39.09$, $P < 0.001$. Magnitudes of the factor loadings were strong (range, 0.36–0.85), and all but two of the 10 correlations among the reverse item residuals were significant (M = 0.09, range, 0.02–0.20).

The same five models were tested in split sample 2 (n = 600). A very similar pattern of results was obtained. Strong correlated residuals were found between items 9 and 16 ($\theta_{16,9}$ = 0.20, $P < 0.001$) and items 7 and 15 ($\theta_{17,5}$ = 0.09, $P < 0.001$) and thus were specified in the hypothesized model (one-factor with method effects) and competing model (two-factors of ‘Worry Engagement’ and ‘Absence of Worry’). As seen in Table 1, the fit of the hypothesized one-factor model was significantly better than that of the competing two-factor solution, $\chi^2_{\text{diff}}(8) = 20.56$, $P < 0.01$. Although somewhat lower than the estimate obtained in the first subsample ($\phi_{2,1} = 0.87$), the factors arising from the two-factor model were highly correlated ($\phi_{2,1} = 0.79$).

To maximize precision of the final parameter estimates, the hypothesized one-factor model was fit to the data of the full sample ($n = 1200$). This model fit the data well, $\chi^2(92) = 532.90$, $P < 0.001$, RMSEA = 0.063 (90% CI = 0.058, 0.068), SRMR = 0.028, CFI = 0.959, TFI = 0.947.[1] Fit diagnostics indicated no salient points of strain in the solution. Factor loadings, which are provided in Table 2, ranged from 0.32 (item 1) to 0.85 (item 7). All error correlations were statistically significant and ranged from 0.05 ($\theta_{10,1}$) to 0.22 ($\theta_{16,9}$); the average of the correlations among reverse-worded items was 0.11.

Although rarely reported in applied psychometric research, evaluation of factor determinacy is an important aspect of factor analytic findings (a highly indeterminate factor can produce radically different factor scores that are nonetheless equally consistent with the obtained factor loadings; cf. Floyd & Widaman, 1995; Grice, 2001). A factor determinacy coefficient (correlation between factor score estimates and the respective factor) was computed for this solution using the Mplus software. As shown in Table 2, a high degree of determinacy was found ($\rho = 0.97$), indicating that PSWQ factor score estimates could serve as suitable substitutes for the factor itself in scenarios where latent structural analyses are not feasible (cf. Grice, 2001).

2.2. Tests of PSWQ invariance in male and females patients

To further evaluate the stability and generalizability of the one-factor solution, the degree of measurement invariance (e.g. equal factor loadings, indicator intercepts) and population homogeneity (e.g. equal factor variances and *M*s) were examined in male and female patients using multiple-groups CFA (*n*s = 462 and 738 for males and females, respectively). Although prior studies have addressed this issue (Brown et al., 1992; Meyer et al., 1990), these analyses were performed in a cursory fashion (*t*-tests of coarse composite scores) in the absence of CFA-based evaluation of the many possible sources of noninvariance (e.g. differing factor structures or factor loadings). After separate CFA models were conducted to ensure adequate fit in the male and female subsamples, a two-group CFA was conducted to simultaneously evaluate equal PSWQ form between sexes. This model fit the data well, $\chi^2(184) = 652.47$, $P < 0.001$, RMSEA = 0.065 (90% CI = 0.060, 0.071), SRMR = 0.031, CFI = 0.957, TFI = 0.944. The factor loadings and factor determinacies for males and females are presented in Table 2. Given evidence of equal form, the parameter equivalence of the PSWQ was evaluated in a series of two-group CFAs that entailed increasingly restrictive constraints (error variances and covariances were freely estimated in all solutions; cf. Byrne, Shavelson & Muthén, 1989). Equality constraints to the factor loadings did not significantly degrade the fit of the model, $\chi^2_{\text{diff}}(15) = 11.15$, *ns*, suggesting that the items

[1] Initial data diagnostics indicated no departures from normality for the PSWQ items except item 10 (‘I never worry about anything’) which evidenced positive skew and leptokurtosis (patients infrequently endorsed this item as typical of them). Analyses were also conducted using robust estimators (robust maximum likelihood, MLM; robust weighted least squares, WLSMV) and no substantial differences in overall model fit or the significance test of parameter estimates were noted; e.g. MLM fit of the hypothesized model with $n = 1200$, $\chi^2(92) = 423.77$, $P < 0.001$, RMSEA = 0.055, SRMR = 0.027, CFI = 0.962, TFI = 0.951.

Table 2
Latent structure of the Penn State Worry Questionnaire: Confirmatory factor analysis using the full sample (n = 1200) and male (n = 462) and female (n = 738) patients

	Item	Full sample	Males	Females
1	If I don't have enough time to do everything, I don't worry about it (R)	0.321	0.394	0.280
2	My worries overwhelm me	0.763	0.763	0.763
3	I do not tend to worry about things (R)	0.569	0.645	0.520
4	Many situations make me worry	0.784	0.806	0.768
5	I know I shouldn't worry about things, but I just cannot help it	0.822	0.818	0.821
6	When I am under pressure I worry a lot	0.673	0.708	0.651
7	I am always worrying about something[a]	0.850	0.846	0.852
8	I find it easy to dismiss worrisome thoughts (R)	0.588	0.589	0.585
9	As soon as I finish one task, I start to worry about everything else I have to do	0.655	0.695	0.630
10	I never worry about anything (R)	0.555	0.589	0.534
11	When there is nothing more I can do about a concern, I don't worry about it any more (R)	0.457	0.477	0.442
12	I've been a worrier all my life	0.616	0.622	0.607
13	I notice that I have been worrying about things	0.773	0.756	0.781
14	Once I start worrying, I can't stop	0.790	0.773	0.799
15	I worry all the time	0.843	0.850	0.838
16	I worry about projects until they are done	0.631	0.634	0.627
	Factor determinacy	0.970	0.971	0.969
	Scale reliability	0.912	0.923	0.905
	Mean	60.74	59.22	61.70
	SD	13.62	13.72	13.48

(R), reverse-scored item; means and S.D.s are based on simple summation of the 16 items.
[a] Item 7 was used as marker variable in all analyses.

measure the factors in comparable ways in males and females. The second analysis indicated that the item intercepts were invariant in the two groups, $\chi^2_{\mathrm{diff}}(15) = 10.48$, *ns*. Similarly, the model which constrained the factor variances to equality did not change the fit of the solution, $\chi^2_{\mathrm{diff}}(1) = 0.05$, *ns*. The final model, which held the latent factor means equal, produced a significant decrease in goodness of fit, $\chi^2_{\mathrm{diff}}(1) = 9.69$, $P < 0.001$. This outcome reflected a sex difference in latent factor scores; specifically, that females scored significantly ($P < 0.005$) higher on the dimension of pathological worry than did males (unstandardized latent M difference = -0.218, $z = 3.11$). Thus, the collective findings suggest a high degree of measurement and structural invariance of the PSWQ in male and female patients, with the exception of a group difference in relative standing on the latent dimension of worry.[2] In addition to the fact that this analysis

[2] Although group differences were not noted until the final step of this evaluation, such analyses may proceed in the context of partial measurement invariance (e.g. a noninvariant pattern of factor loadings) so long as the model specification entails at least one indicator loading, intercept, etc., that is invariant (Byrne, Shavelson & Muthén, 1989; Muthén & Christoffersson, 1981).

was highly powered to detect small effects as statistically significant (cf. minimal differences in latent factor and coarse composite score *M*s), this difference could be due in part to a slight, but significant, sex difference in current GAD (19 and 24% of males and females had current GAD, respectively), $\chi^2(1) = 4.07$, $P < 0.05$, four-fold point correlation (V) = 0.058.

2.3. Scale reliabilities

The scale reliability of the PSWQ was estimated within the hypothesized one-factor CFA model using the approach developed by Raykov (2001). This method reconciles the problems with Cronbach's α which is a misestimator of reliability except in the rare instance when all elements of a multiple-item measure are tau-equivalent and all measurement error is random (Lord & Novick, 1968; McDonald, 1999; Raykov, 2001). In LISREL, the procedure entails specifying three dummy latent variables whose variances are constrained to equal the numerator (true score variance), denominator (total variance), and corresponding ratio of true score variance to total score variance, per the classic formula for scale reliability estimation (Lord & Novick, 1968). As shown in Table 2, the scale reliability was quite favorable for the full sample (ρ = 0.912), and for male and female patients (ρs = 0.923 and 0.905, respectively).

An additional advantage of this CFA-based approach is that scale reliabilities can be estimated within the context of a specified error structure. In the present case, failure to model the error covariances may result in an over-estimation of scale reliability because some of the item covariance due to method effects would be absorbed by the factor loadings. To illustrate, the PSWQ scale reliabilities were re-estimated using one-factor models where all indicator residual variances were orthogonal. The analyses resulted in scale reliabilities of 0.933, 0.938, and 0.929 for the full sample, males, and females, respectively. Thus, when an error structure was ignored, estimates of PSWQ true score variance were 1.6 to 2.6% higher than estimates from the one-factor solution which incorporated error covariances reflecting method effects.

Finally, the scale reliabilities were re-estimated using only the 11 items worded in the symptomatic direction (with error covariances between items 7 and 15, and items 9 and 16). The analyses were performed on the basis of the suggestion that reverse-worded items be dropped in scale scoring in many empirical applications to avoid complications from incorporating a complex error structure (Marsh, 1996). These solutions yielded scale reliability estimates of 0.921 for the full sample, and 0.924 and 0.918 for male and female patients, respectively.

3. Discussion

The present findings provide strong and consistent evidence (across independent subsamples, male and female patients) that the covariances among the 16 items of the PSWQ are best explained by a single underlying construct (i.e. 'excessive/uncontrollable worry,' as supported by validity analyses in several prior investigations; e.g. Brown et al., 1992, 1998; Meyer et al., 1990). Although refuting previous conclusions that the latent structure of the PSWQ consists of two potentially meaningful factors ('Worry Engagement,' 'Absence of Worry'; e.g. Beck et al., 1995; Fresco et al., 2002), these results are consistent with an extensive literature attesting to the presence of method effects in scales entailing a combination of positively and negatively worded items (e.g. Bagozzi & Heatherton, 1994; Marsh, 1996). As many researchers have noted (e.g. Marsh, 1996; Tomás & Oliver, 1999), if these effects are not modeled in factor analyses of such instruments, the resulting latent structure of the scale responses may be confounded or obscured by method bias.

Because reverse-worded items are often associated with method effects, their inclusion complicates the interpretation and psychometric evaluation of psychological tests. Whereas psychometric tradition encourages the use of both positively and negatively worded items in scale development, Marsh (1996) has argued that the advantages of this approach may be often outweighed by these complications. While noting that the easiest way to avoid such problems is to design scales without reverse-worded items, Marsh (1996) recommended that when positively and negatively worded items are used, their proportion should be equal ('without this balance, it is difficult to establish how much the distinction between different factors is due to differences in the underlying constructs being measured as opposed to method effects'; p. 817).

This point is germane to the psychometric evaluation of the PSWQ, because only five of its 16 items are reverse worded. For instance, recent structural analyses of Rosenberg's (1965) SES questionnaire have examined the comparative fit of measurement models in which the error covariances of either the positively worded or negatively worded items were freely estimated (e.g., Marsh, 1996; Tomás & Oliver, 1999; Wang, Siegal, Falck & Carlson, 2001).[3] These model comparisons are performed to determine if method effects are more prominent in positively or negatively worded items. Although relatively straightforward using the SES (five positively worded, five negatively worded items), such analysis with the PSWQ is confounded because non-reversed statements outnumber reversed items by a ratio of over 2:1. Indeed, this marked imbalance precludes meaningful nested model evaluation; i.e. it would artifactually favor the model specifying correlated residuals among the 11 items worded in the symptomatic direction because 55 covariances are freely estimated vs. 10 covariances in the reverse-item counterpart model.

As an alternative strategy, Marsh (1996) has suggested that a small number of reverse-worded items could be included in psychological scales to disrupt or evaluate potential response biases, but not used in test scoring. Because most studies will use the PSWQ without accommodation of a complex error structure, this suggestion merits consideration in the future scoring of this measure. Fresco et al. (2002) reported that PSWQ scores based solely on the 11 positively worded items were equally or more strongly correlated to convergent validity indices relative to the full 16-item measure, although these findings were based on coarse composites (summation of raw scores) rather than factor scores or CFA-based estimates that incorporated an appropriate error theory. In the present study, scale reliability estimates for the 11 items were slightly higher than the estimates for the full PSWQ scale (e.g. 0.921 vs 0.912 for the 11- and 16-item versions in the full sample). In accord with the conclusions of Fresco et al. (2002), these preliminary results indicate the potential viability of limiting PSWQ scale scoring to 11 items, although further CFA-based validation is needed.

Similar to Marsh (1996), the present study's analyses represent *correlated uniqueness* models frequently used in CFA research of multitrait-multimethod (MTMM) data (e.g. Byrne & Goffin,

[3] Models where the correlated uniquenesses of positively and negatively worded items are *simultaneously* estimated are typically not reported because of improper solutions (Marsh, 1996).

1993; Marsh & Grayson, 1995; Tomás, Hontangas & Oliver, 2000). It is noteworthy that MTMM and psychological test data (e.g. the SES) have also been evaluated using *correlated methods* approaches. Unlike correlated uniqueness models where error covariances are freely estimated, correlated methods models involve specification of latent factors reflecting method effects, in addition to the factor representing the substantive underlying construct (with the PSWQ, this would entail an 'excessive/uncontrollable worry' factor on which all 16 items would load, and two separate method factors for the 11 positively and five negatively worded items). Many researchers have observed that the correlated method approach has some strengths (e.g. addresses the interpretative nature of the method effects and allows these effects to be intercorrelated; is often a more parsimonious solution than correlated uniqueness models), but these advantages are offset by several limitations including frequent improper solutions (Heywood cases, convergence problems) and the inability to model method effects in a multidimensional fashion (e.g. Bagozzi, 1993; Kumar & Dillon, 1992; Marsh & Grayson, 1995; Tomás & Oliver, 1999). Although the correlated uniqueness approach has some limitations (e.g. assumes method effects are uncorrelated), it is generally favored by methodologists due to its advantages of rarely producing improper solutions and not adhering to the assumption that methods are unidimensional (e.g. the error covariance of some PSWQ items may be due to both the direction and content of wording; cf. Marsh & Bailey, 1991; Marsh & Grayson, 1995).

The present study's findings underscore the importance of considering the potential existence of method effects of any type (e.g. wording direction, high content overlap) in the evaluation of CFA measurement models. Despite the fact that nonrandom measurement error is common in multiple-item scales (cf. Byrne, Shavelson & Muthén, 1989; Floyd & Widaman, 1995; Gerbing & Anderson, 1984), applied researchers often fail to report whether such effects were present in their solutions (i.e. limit model evaluation to indices of overall fit), but instead rely on the CFA only to compare the fit of competing models (e.g. one- vs two-factor solutions). Although researchers can often forward an a priori error theory (e.g. error covariance due to positively and negatively worded items), the presence of correlated residuals is frequently revealed after initial model specification (e.g. PSWQ items 7 and 15). Although post hoc model fitting is often criticized (e.g. MacCallum, 1986), other researchers (e.g. Byrne, Shavelson & Muthén, 1989) assert that this process can be substantively meaningful and can eliminate salient biases in the estimation of the model parameters of interest (e.g. factor loadings, scale reliabilities; cf. Green & Hershberger, 2000; Raykov, 2001). Of course, this process should not be undertaken solely to improve model fit, but must be guided by substantive justification, and ideally, be supported by independent sample replication (cf. split sample 1 and 2 results).

Although underutilized in the applied psychometric literature, the procedure of 'exploratory factor analysis within CFA' (E/CFA; Jöreskog, 1969) can be a useful precursor to CFA that allows the researcher to explore measurement structures more fully before moving into a confirmatory framework.[4] A common sequence in psychological scale development is to conduct CFA as the next step after a measure's latent structure has been explored using EFA. However, the researcher will frequently encounter a poor-fitting CFA solution because of the multiple potential sources of misfit that are not present in EFA (e.g. unlike EFA, indicator cross-loadings and residual

[4] I am grateful to Bengt and Linda Muthén for alerting me to this approach.

covariances are often fixed to zero in initial CFA models). The researcher is then faced with potentially extensive post hoc model testing subject to the criticisms of specification searches in a single data set (MacCallum, 1986). The E/CFA approach represents an intermediate step between EFA and CFA that provides substantial information important in the development of realistic confirmatory solutions. In this strategy, the CFA applies the same number of identifying restrictions used in EFA (m^2) by fixing factor variances to unity and by selecting one anchor item for each factor whose cross-loadings are fixed to zero (the loadings of non-anchor items are freely estimated on each factor). Whereas this specification produces the same model fit as maximum likelihood EFA, the CFA estimation provides considerably more information including the significance of cross-loadings and the presence of salient error covariances (method effects). Thus, the researcher can develop a realistic measurement structure prior to moving into a more restrictive CFA framework (E/CFA was not used in the current study due to the strong a priori basis for a unidimensional structure; for an applied example of this approach using a multifactorial measurement model, see Brown, White, Forsyth, & Barlow, in press).

While the present findings provide compelling evidence that a single, substantively meaningful construct underlies the responses of the 16 PSWQ items, additional work could be undertaken to further explore the psychometric behavior of this instrument. As noted earlier, subsequent research may support the notion that the five reverse-worded items do not contribute meaningfully to the PSWQ measurement properties (e.g. scale reliability, concurrent and predictive validity) beyond the evaluation or possible disruption of response bias (although, to date, these latter issues have not been well-explored in empirical and clinical applications of this measure). Although the current study's findings indicated that the PSWQ measurement properties were invariant in male and female patients, this line of inquiry could be meaningfully expanded to examine the degree of invariance in other subpopulations (e.g. clinical vs. nonclinical samples; comparisons across racial, ethnic, and other demographic groups). Along these lines, it is noteworthy that the type of method effect addressed in this study may be noninvariant across relevant subgroups. In his studies with students grade 8 and below, Marsh (1986, 1996) has shown that method effects associated with negatively worded SES items diminish as a function of children's verbal ability. It would be of interest to determine whether such method effects in clinical instruments are moderated by salient dimensions or subgroups (e.g. less evident in nonclinical vs clinical samples; in treated clinical samples, less evident at post-treatment as a function of decreased acute distress).

The issues addressed in this paper were discussed in specific context of the PSWQ, but have broad applicability to the psychometric evaluation of psychological tests of all types (e.g. method effects/MTMM models, factor determinacy, CFA-based scale reliability, E/CFA, measurement invariance of mean and covariance structures). Indeed, the CFA framework offers many underutilized opportunities and advantages to the applied psychometric researcher.

Acknowledgements

The author thanks Laura Campbell, Jessica Grisham and Melanie Santos for their assistance with this study.

References

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. rev.). Washington, DC: Author.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bachman, J. G., & O'Malley, P. M. (1986). Self-concepts, self-esteem, and educational experiences: The frog pond revisited (again). *Journal of Personality and Social Psychology*, *50*, 35–46.

Bagozzi, R. P. (1993). An examination of the psychometric properties of measures of negative affect in the PANAS-X scales. *Journal of Personality and Social Psychology*, *65*, 836–851.

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling*, *1*, 35–67.

Barlow, D. H., Rapee, R. M., & Brown, T. A. (1992). Behavioral treatment of generalized anxiety disorder. *Behavior Therapy*, *23*, 551–570.

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*, 143–156.

Beck, J. G., Stanley, M. A., & Zebb, B. J. (1995). Psychometric properties of the Penn State Worry Questionnaire in older adults. *Journal of Clinical Gerontology*, *1*, 33–42.

Beck, J. G., Stanley, M. A., & Zebb, B. J. (1996). Characteristics of generalized anxiety disorder in older adults: A descriptive study. *Behaviour Research and Therapy*, *34*, 225–234.

Borkovec, T. D., & Costello, E. (1993). Efficacy of applied relaxation and cognitive behavioral therapy in the treatment of generalized anxiety disorder. *Journal of Consulting and Clinical Psychology*, *61*, 611–619.

Brown, T. A., Antony, M. M., & Barlow, D. H. (1992). Psychometric properties of the Penn State Worry Questionnaire in a clinical anxiety disorders sample. *Behaviour Research and Therapy*, *30*, 33–37.

Brown, T. A., Barlow, D. H., & Liebowitz, M. R. (1994). The empirical basis of generalized anxiety disorder. *American Journal of Psychiatry*, *151*, 1272–1280.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the *DSM-IV* anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology*, *107*, 179–192.

Brown, T. A., Di Nardo, P. A., Lehman, C. L., & Campbell, L. A. (2001). Reliability of *DSM-IV* anxiety and mood disorders: Implications for the classification of emotional disorders. *Journal of Abnormal Psychology*, *110*, 49–58.

Brown, T. A., Marten, P. A., & Barlow, D. H. (1995). Discriminant validity of the symptoms constituting the *DSM-III-R* and *DSM-IV* associated symptom criterion of generalized anxiety disorder. *Journal of Anxiety Disorders*, *9*, 317–328.

Brown, T. A., Moras, K., Zinbarg, R. E., & Barlow, D. H. (1993). Diagnostic and symptom distinguishability of generalized anxiety disorder and obsessive-compulsive disorder. *Behavior Therapy*, *24*, 227–240.

Brown, T. A., White, K. S., Forsyth, J. P., & Barlow, D. H. (in press). The structure of perceived emotional control: Psychometric properties of a revised Anxiety Control Questionnaire. *Behavior Therapy*.

Butler, G., Wells, A., & Dewick, H. (1995). Differential effects of worry and imagery after exposure to a stressful stimulus: A pilot study. *Behavioural and Cognitive Psychotherapy*, *23*, 45–56.

Byrne, B., & Goffin, R. (1993). Modeling multitrait-multimethod data from additive and multiplicative covariance structures: An audit of construct validity concordance. *Multivariate Behavioral Research*, *28*, 67–96.

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.

Byrne, D., Fisher, D., Lamberth, J., & Mitchell, H. E. (1974). Evaluation of erotica: Facts or feelings. *Journal of Personality and Social Psychology*, *29*, 111–116.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Coles, M. E., Mennin, D. S., & Heimberg, R. G. (2001). Distinguishing obsessive features and worries: The role of thought-action fusion. *Behaviour Research and Therapy*, *39*, 947–959.

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised job diagnostic survey: Is education a factor in responses to negatively worded items? *Journal of Applied Psychology*, *78*, 141–143.

Di Nardo, P. A., Brown, T. A., & Barlow, D. H. (1994). *Anxiety Disorders Interview Schedule for DSM-IV: Lifetime version (ADIS-IV-L)*. San Antonio, TX: Psychological Corporation.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286–299.

Fresco, D. M., Heimberg, R. G., Mennin, D. S., & Turk, C. L. (2002). Confirmatory factor analysis of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *40*, 313–323.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research*, *11*, 572–580.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling*, *7*, 251–270.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430–450.

Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.51 [Computer software]*. Chicago: Scientific Software International.

Kumar, A., & Dillon, W. R. (1992). An integrative look at the use of additive and multiplicative covariance models in the analysis of MTMM data. *Journal of Marketing Research*, *29*, 51–64.

Ladouceur, R., Dugas, M. J., Freeston, M. H., Léger, E., Gagnon, F., & Thibodeau, N. (2000). Efficacy of a cognitive-behavioral treatment for generalized anxiety disorder: Evaluation in a controlled clinical trial. *Journal of Consulting and Clinical Psychology*, *68*, 957–964.

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

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107–120.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. *Developmental Psychology*, *22*, 37–49.

Marsh, H. W. (1992). *Self-Description Questionnaire III: Manual*. Macarthur, Australia: Faculty of Education, University of Western Sydney.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810–819.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analysis of multitrait-multimethod data: A comparison of the behavior of alternative models. *Applied Psychological Measurement*, *15*, 47–70.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousand Oaks, CA: Sage.

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

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn state worry questionnaire. *Behaviour Research and Therapy*, *28*, 487–495.

Molina, S., & Borkovec, T. D. (1994). The Penn State Worry Questionnaire: Psychometric properties and associated characteristics. In G. C. L. Davey, & F. Tallis (Eds.), *Worrying: Perspectives on theory, assessment and treatment* (pp. 265–283). New York: Wiley & Sons.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous indicators in several groups. *Psychometrika*, *46*, 407–419.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus 2.02 [Computer software]*. Los Angeles: Author.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*, 315–323.

Rosenberg, M. (1965). *Society and the adolescent child*. Princeton, NJ: Princeton University Press.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The results of careless respondents? *Applied Psychological Measurement*, *9*, 367–373.
Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artificial factors. *Journal of Management*, *23*, 659–677.
Starcevic, V. (1995). Pathological worry in major depression: A preliminary report. *Behaviour Research and Therapy*, *33*, 55–56.
Stöber, J. (1995). Besorgnis: Ein Vergleich dreier Inventare zur Erfassung allgemeiner Sorgen [Worry: A comparison of three questionnaires for the measurement of general worries]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *16*, 50–63.
Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.
Tomás, J. M., Hontangas, P. M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research*, *35*, 469–499.
Tomás, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling*, *6*, 84–98.
van Rijsoort, S., Emmelkamp, P., & Vervaeke, G. (1999). The Penn State Worry Questionnaire and the Worry Domains Questionnaire: Structure, reliability, and validity. *Clinical Psychology and Psychotherapy*, *6*, 297–307.
Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001). Factorial structure of Rosenberg's self-esteem scale among crack-cocaine drug users. *Structural Equation Modeling*, *8*, 275–286.
Wells, A., & Papageorgiou, C. (1998). Relationships between worry, obsessive-compulsive symptoms, and meta-cognitive beliefs. *Behaviour Research and Therapy*, *36*, 899–913.
Widaman, K. F. (1993). Common factor analysis versus principal components analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263–311.