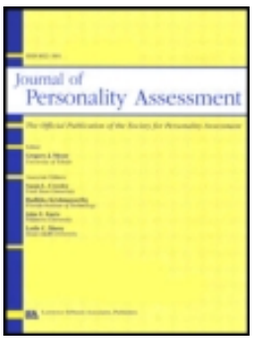

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

Scoring and Modeling Psychological Measures in the Presence of Multidimensionality

Steven P. Reise , Wes E. Bonifay & Mark G. Haviland

To cite this article: Steven P. Reise , Wes E. Bonifay & Mark G. Haviland (2013) Scoring and Modeling Psychological Measures in the Presence of Multidimensionality, Journal of Personality Assessment, 95:2, 129-140, DOI: 10.1080/00223891.2012.725437

To link to this article: https://doi.org/10.1080/00223891.2012.725437

Journal of Personality Assessment, *95*(2), 129–140, 2013

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223891.2012.725437

Routledge
Taylor & Francis Group

STATISTICAL DEVELOPMENTS AND APPLICATIONS

Scoring and Modeling Psychological Measures in the Presence of Multidimensionality

STEVEN P. REISE,[1] WES E. BONIFAY,[1] AND MARK G. HAVILAND[2]

[1]*Department of Psychology, University of California, Los Angeles*
[2]*Department of Psychiatry, Loma Linda University*

Confirmatory factor analytic studies of psychological measures showing item responses to be multidimensional do not provide sufficient guidance for applied work. Demonstrating that item response data are multifactorial in this way does not necessarily (a) mean that a total scale score is an inadequate indicator of the intended construct, (b) demand creating and scoring subscales, or (c) require specifying a multidimensional measurement model in research using structural equation modeling (SEM). To better inform these important decisions, more fine-grained psychometric analyses are necessary. We describe 3 established, but seldom used, psychometric approaches that address 4 distinct questions: (a) To what degree do total scale scores reflect reliable variation on a single construct? (b) Is the scoring and reporting of subscale scores justified? (c) If justified, how much reliable variance do subscale scores provide after controlling for a general factor? and (d) Can multidimensional item response data be represented by a unidimensional measurement model in SEM, or are multidimensional measurement models (e.g., second-order, bifactor) necessary to achieve unbiased structural coefficients? In the discussion, we provide guidance for applied researchers on how best to interpret the results from applying these methods and review their limitations.

Many personality and psychopathology measures are designed to assess individual differences on a single target construct; yet, the pages of assessment journals are filled with confirmatory factor analytic (CFA) work that repeatedly demonstrate item responses to the measure of interest having a multidimensional structure (i.e., multiple correlated dimensions account for the common variance better than a unidimensional model). A major conclusion commonly drawn from CFA research is that multidimensional structures support the development and scoring of subscales and, in fact, call into question the use of total scale scores. Findings of multidimensionality also suggest that if item responses to a measure are used to represent a latent variable in structural equation modeling (SEM), a multidimensional measurement model (e.g., second-order, bifactor, correlated factors) is required.

For one notable example, the multidimensionality of the Beck Depression Inventory II (BDI–II; Beck, Steer, & Brown, 1996) has received extensive CFA scrutiny (recent examples summarized and critiqued in Brouwer, Meijer, & Zevalkink [2012] include Al-Turkait and Ohaeri, 2010; Osman, Barrios, Gutierrez, Williams, & Bailey, 2008; Quilty, Zhang, & Bagby, 2010; Vanheule, Desmet, Groenvynck, Rosseel, & Fontaine, 2008; Ward, 2006). Although there is some consistency across all studies (e.g., item response data are not strictly unidimensional), debates about the appropriateness of one model over another, how many factors, and recommendations about scoring vary considerably from one study to the next. The BDI–II is by no means an exception; indeed, the dimensionality of item responses on most major clinical and personality assessment scales has been studied extensively across various college student, community, and clinical samples, and similar conclusions have been drawn. In the case of the BDI–II, Brouwer et al. (2012), after their thorough evaluations, including bifactor model application, reached a clear and compelling conclusion that BDI–II total scores are good estimates of overall depression severity and that subscale scores present too many interpretive challenges.

CFA findings of multidimensionality make an important contribution toward our understanding of the systematic influences that affect item responses, and such findings *potentially* have important applied consequences in terms of scale scoring and interpretation. For example, if item response data are multidimensional:

1. Item subsets (subscales) (a) are not interchangeable indicators of a single construct, (b) might relate differently to an external criterion (e.g., fMRI results), and (c) could have distinct implications for health policy and psychological intervention.
2. Total scale scores might not reflect the target construct of interest because their interpretation could be confounded hopelessly by multiple systematic sources of variance.
3. In an SEM context, a researcher might need to consider the use of complicated, multidimensional latent variable model specifications (e.g., second-order or bifactor measurement models; see Chen, West, & Sousa, 2006; Thomas, 2012).

Unfortunately, most CFA investigations of a measure's factor structure do not directly show whether these potential consequences of multidimensionality are relevant to a particular measure and often raise as many applied questions as they answer.

Received May 28, 2010; Revised June 23, 2012.

Address correspondence to Steven P. Reise, Department of Psychology, University of California, Los Angeles, Franz Hall, Los Angeles, CA 90095; Email: reise@psych.ucla.edu

CFA investigations, in fact, seldom inform researchers on the degree to which multidimensionality is severe enough such that the total score is uninterpretable as an indicator of a single construct or whether subscales are psychometrically justified or how much reliable information a subscale score provides beyond the total score.

Without knowledge of these issues or roadmaps in the psychometric and clinical literatures for answering the questions they raise in applied work, decisions about how to proceed in the presence of multidimensionality appear to be arbitrary. The main objective of this study, thus, is to review tools that researchers and clinicians can use for systematically evaluating the effects of multidimensionality in their work. In what follows, we apply three modern statistical methods to demonstrate how researchers can study the effects of multidimensionality to address the following four distinct questions:

1. To what degree do total scale scores reflect reliable variation on a single construct?
2. Is the scoring and reporting of subscale scores justified?
3. If justified, how much reliable variance do subscale scores provide after controlling for a general factor?
4. Can multidimensional item response data be represented by a unidimensional measurement model in research with SEM, or are multidimensional measurement models (e.g., second-order, bifactor) necessary to achieve unbiased structural coefficients?

Although our demonstrations are relevant to a wide range of personality, psychiatric, medical, and clinical assessment instruments, we use only one as a running example throughout, the Twenty-Item Toronto Alexithymia Scale (TAS–20; Bagby, Parker, & Taylor, 1994). We selected the TAS–20 for three reasons: (a) the availability of a relatively large data set, (b) our long-standing interest in the structure of the TAS–20 (Haviland & Reise, 1996b) and the definition and measurement of alexithymia (Haviland & Reise, 1996a), and most important, (c) the lack of clear guidance or justification in the alexithymia and psychometric literatures for TAS–20 scoring (calculating total or subscale scores or both) and use in SEM.

The Multidimensional Structure of the TAS–20

For many measures, the publication of competing CFAs arguing for different multidimensional structures has become a cottage industry, and the TAS–20 (Bagby et al., 1994) is no exception. In brief, the TAS–20 is an influential self-report scale used in hundreds of studies of the experience and regulation of affect. There are many CFAs of TAS–20 data across an array of people in general and clinical samples worldwide. Although there are variations in the findings, and debates about the "true" structure of the instrument continue (e.g., Bagby, Taylor, Quilty, & Parker, 2007; Gignac, Palmer, & Stough, 2007; Haviland & Reise, 1996b; Meganck, Vanheule, & Desmet, 2008; Müller, Bühner, & Ellgring, 2003; Taylor, Bagby, & Luminet, 2000), in almost all studies, investigators reach the same conclusion, namely, data generated from the TAS–20 do not fit a unidimensional model, and, thus, have a multidimensional structure (commonly, debates center around two vs. three or four factors).

Table 1.—Twenty-Item Toronto Alexithymia Scale (TAS–20) item numbers, subscales, and content.

1	DIF	I am often confused about what emotion I am feeling.
3	DIF	I have physical sensations that even doctors don't understand.
6	DIF	When I am upset I don't know if I am sad, frightened, or angry.
7	DIF	I am often puzzled by sensations in my body.
9	DIF	I have feelings that I can't quite identify.
13	DIF	I don't know what's going on inside me.
14	DIF	I often don't know why I am angry.
2	DDF	It is difficult for me to find the right words for my feelings.
4	DDF	I am able to describe my feelings easily.
11	DDF	I find it hard to describe how I feel about people.
12	DDF	People tell me to describe my feelings more.
17	DDF	It is difficult for me to reveal my innermost feelings, even to close friends.
5	EOT	I prefer to analyze problems rather than just describe them.
8	EOT	I prefer to just let things happen rather than to understand why they turned out that way.
10	EOT	Being in touch with emotions is essential.
15	EOT	I prefer talking to people about their daily activities rather than their feelings.
16	EOT	I prefer to watch "light" entertainment shows rather than psychological dramas.
18	EOT	I can feel close to someone, even in moments of silence.
19	EOT	I find examination of my feelings useful in solving problems.
20	EOT	Looking for hidden meanings in movies or plays distracts from their enjoyment.

Note. DIF = difficulty identifying feelings; DDF = difficulty describing feelings; EOT = externally oriented thinking.

Perhaps the most usual finding, however, and the structure endorsed by the scale authors (Parker, Taylor, & Bagby, 2003; Taylor, Bagby, & Parker, 2003) is that the correlations among the TAS–20 items can be explained by three correlated latent factors: (a) difficulty identifying feelings (DIF), (b) difficulty describing feelings (DDF), and (c) externally oriented thinking (EOT). Item content and the assignment of items to these three factors are shown in Table 1. This three-factor finding is so widely accepted that applied researchers routinely calculate and use both total scale score and three subscale scores (examples abound), despite the test authors' recommendation to use total scores only as a continuous measure of alexithymia severity (see Parker, Keefer, Taylor, & Bagby, 2008). As such, the literature is replete with studies reporting statistically significant relationships between total TAS–20 scores and an external correlate and relationships between one, two, or three subscale scores and the correlate of interest (or not with total score but with one or two subscales, and so forth; see Lumley, Neely, & Burger, 2007). Use of the TAS–20 in SEM is equally inconsistent, with researchers sometimes using all items and sometimes using items from only one or two subscales (typically DIF or DIF and DDF; e.g., Hund & Espelage, 2005; Ouwens, van Strien, & van Leeuwe, 2009).

In this study, our goals do not include challenging the results of the various TAS–20 structural analyses, for each has its strengths and limits. We, however, do question how researchers have responded to TAS–20 data being multidimensional; in other words, how they have used such findings in substantive research (e.g., scoring both total and subscale scores) or suggesting that bifactor models be specified in SEM (see Gignac et al., 2007). Later, with both new and published TAS–20 data, we evaluate (a) the interpretability of the total TAS–20 score as reflecting a single construct, (b) whether the scoring

of the three generally agreed on subscales—DIF, DDF, and EOT—can be psychometrically justified, (c) the degree to which scores from the three subscales are reliable after controlling for alexithymia, and (d) the effects of multidimensionality when specifying TAS–20 items as a single latent variable in SEM research. The first of the proposed methods is based on traditional classical test theory, whereas the second two depend on the adoption of a latent variable modeling framework. We begin first by considering a classical test theory approach to judging whether scoring and reporting subscale scores can be justified.

Method 1: Should Subscale Scores Be Reported?

Findings of multidimensionality often are taken as both necessary and sufficient justification for the scoring and reporting of subscale scores. There are several important concerns, however, with this standard practice. In this section, we consider the problem that subscales, by definition, typically have lower (internal consistency) reliability than the total test score. This differential reliability, when combined with the fact that total and subscale scores are correlated, can create conditions where it is possible that the total score is a better indicator of the subscale construct than the subscale score. Fortunately, identifying when these subscale invalidating conditions occur is straightforward, as described later.

Partly due to legislation that mandates the reporting of both total and subscale scores across a broad range of achievement domains, psychometric research on the merits of reporting subscale score profiles has been an important topic in educational assessment as reviewed in Sinharay, Puhan, and Haberman (2011). As many educational assessment researchers have noted, subscale scores often are highly intercorrelated despite earnest attempts to assess examinee competencies on distinct content domains (e.g., algebra, reading comprehension) or cognitive processes. Moreover, it is typical in education that the total score is based on many items and, thus, relatively reliable, whereas subscale scores often are based on fewer items and relatively less reliable.

Haberman (2008) pointed out that under such conditions, it is possible that the aggregate total test score is a better predictor of an individual's true score on a subscale than is the observed subscale score. When this counterintuitive and disturbing result is true, there can be no psychometric justification for reporting a subscale score. To understand Haberman's approach to determining the appropriateness of reporting subscale scores, consider an observed subscale score on, say, (SUB_X). This score can be thought of as an estimate of an individual's true score on the subscale, (SUB_T). In the language of Haberman and others, the mean square error of using observed subscale scores to estimate true subscale scores can be referred to as the proportional reduction in mean squared error based on subscale scores ($PRMSE_S$). This term can be estimated through the computation of coefficient alpha for the subscale. $PRMSE_S$ estimates the degree to which the measurement error on a subscale is reduced based on the subscale reliability.

To the degree that subscales are intercorrelated, however (and, thus, redundant and correlated with the total score), and to the degree that the total score is more reliable than a subscale score, it is possible that the total score (e.g., TOT_X) is a better predictor of (SUB_T) than (SUB_X). If so, then the subscale score is a less precise indicator of relative standing on the subscale construct than the total score. As a consequence, it is arguable that the subscale score provides no "added value" beyond the total score and, thus, should not be computed, reported, or used to make educational or clinical decisions. There would be no point, for example, in reporting a separate reading comprehension subscale score if, in fact, the aggregate total score were a better predictor of true score standing on reading comprehension than the reading comprehension subscale score.

To determine whether a total score is a better estimator of subscale true scores than the subscale score, one needs to compute the proportional reduction in mean squared error based on total scores ($PRMSE_{TOT}$) and compare it with $PRMSE_S$. Haberman (2008) provided a set of simple equations to estimate $PRMSE_{TOT}$. Required are only the reliabilities of the total and subscale scores, standard deviations of the total and subscale scores, and intercorrelation matrix among the subscale scores—materials that should be provided in a psychometric report.

In what follows, we illustrate the Haberman procedure based on TAS–20 data ($N = 1{,}612$) collected from a college student sample. The data were provided by Tuppett M. Yates, PhD (University of California–Riverside, Department of Psychology), and, thus, we have named them the Yates data. In the Yates data, coefficient alpha reliabilities for total and subscale scores are .85 for TAS–20, .86 for DIF, .77 for DDF, and, .60 for EOT. The standard deviations are 11.51 for TAS–20, 5.89 for DIF, 4.34 for DDF, and 4.37 for EOT, and the subscale intercorrelations are .63 (DIF and DDF), .27 (DIF and EOT), and .35 (DDF and EOT).[1]

1. The first step is to compute the true score variance for the total and each subscale. These simply are the observed score variance multiplied by the reliability estimate.

$$\text{VAR(true)} = \text{VAR(observed)} \times \text{reliability estimate}$$

$$\text{VAR}(\text{DIF}_T) = 5.89^2 \times .86 = 29.90$$

$$\text{VAR}(\text{DDF}_T) = 4.34^2 \times .77 = 14.53$$

$$\text{VAR}(\text{EOT}_T) = 4.37^2 \times .60 = 11.47$$

$$\text{VAR}(\text{TAS}_T) = 11.51^2 \times .85 = 112.54$$

2. We then compute the covariance matrix among *true* subscale scores. In this case, this will be a 3×3 matrix. On the diagonal are the true subscale variance estimates calculated earlier. On the off-diagonal are covariances between true scores for subscale pairs. Note that these simply are the covariances of the observed subscale scores because a covariance between true scores is equivalent to the covariance between

[1]The numbers provided in the tables and equations have been rounded to fewer places than those used in the computer-generated calculations, and there is not perfect correspondence between hand and computer calculations. The results in the equations and text are the correct figures.

observed scores.

$$\begin{matrix} \mathrm{VAR}(\mathrm{DIF_T}) & r(\mathrm{DIF, DDT}) \times \mathrm{SD_{DIF_X}} \times \mathrm{SD_{DDF_X}} & r(\mathrm{DIF, EOT}) \times \mathrm{SD_{DIF_X}} \times \mathrm{SD_{EOT}} \\ r(\mathrm{DDF, DIF}) \times \mathrm{SD_{DDF_X}} \times \mathrm{SD_{DIF_X}} & \mathrm{VAR}(\mathrm{DDF_T}) & r(\mathrm{DDF, EOT}) \times \mathrm{SD_{DDF_X}} \times \mathrm{SD_{EOT_X}} \\ r(\mathrm{EOT, DIF}) \times \mathrm{SD_{EOT_X}} \times \mathrm{SD_{DDF_X}} & r(\mathrm{EOT, DDF}) \times \mathrm{SD_{EOT_X}} \times \mathrm{SD_{DDF_X}} & \mathrm{VAR}(\mathrm{EOT_T}) \end{matrix}$$

$$= \begin{matrix} 29.90 & 16.18 & 7.06 \\ 16.18 & 14.53 & 6.57 \\ 7.06 & 6.57 & 11.47 \end{matrix}$$

3. Summing across row elements yields the covariance between total true scores ($\mathrm{TOT_T}$) and subscale true scores ($\mathrm{S_T}$). Using the present notation, these values can be referred to as:

$$\mathrm{Cov}(\mathrm{SUB_T}, \mathrm{TOT_T}) \quad (1)$$

In the Yates data, these values are 53.15, 37.29, and 25.11, for DIF, DDF, and EOT, respectively. The next to last step is, for each subscale, to convert the covariances in Equation 1 to correlations squared (reliabilities), as in Equation 2:

$$\rho^2(\mathrm{SUB_T}, \mathrm{TOT_T}) = \frac{[\mathrm{Cov}(\mathrm{SUB_T}, \mathrm{TOT_T})]^2}{\mathrm{Var}(\mathrm{SUB_T})\mathrm{Var}(\mathrm{TOT_T})} =$$

$$\frac{[53.15]^2}{(29.90)(112.54)}$$

$$\frac{[37.29]^2}{(14.53)(112.54)}$$

$$\frac{[25.11]^2}{(11.47)(112.54)} \quad (2)$$

which equals .84, .85, and .49 respectively, for DIF, DDT, and EOT.

4. The last step is to convert the values in Equation 2 into correlations squared between total test scores ($\mathrm{TOT_X}$) and true subscale scores ($\mathrm{SUB_T}$). Stated differently, to convert them to $\mathrm{PRMSE_{TAS}}$ values, the results of Equation 2 need to be multiplied by the reliability of the total score (.85), which yields $\mathrm{PRMSE_{TOT}} = .71, .72$, and .42 for DIF, DDF, and EOT, respectively. These values then are compared to $\mathrm{PRMSE_S}$: .86 for DIF, .77 for DDF, and, .60 for EOT. To the degree that the latter are larger than the former, subscale scores provide a relatively better indicator of subscale true score standing, and, thus, can be reported. In this case, subscale scores, indeed, are better measures of subscale true scores relative to the total score, and these results allow one to argue that it is statistically justifiable to report subscale scores for the present TAS–20 data and especially for EOT.

The preceding findings appear not to be unique to the Yates data. Published data from four samples yield similar results (see Table 2); for example, in Zhu et al.'s (2007) student sample ($N = 870$), coefficient alphas were .77, .65, and .62, whereas $\mathrm{PRMSE_{TAS}}$ values were .69, .67, and .38. Similar findings were observed in their clinical sample ($N = 179$), as shown in Table 2. Thus, in these studies, DIF and EOT pass the Haberman subscale scoring test, but DDF does not. From a second study, Culhane, Morera, Watson, and Millsap (2009) reported the necessary statistics for the TAS–20 in both Anglo and Hispanic student samples. As shown in Table 2, these numbers are consistent with the Yates data in supporting the argument that subscale observed scores better reflect subscale true scores, relative to the total scale score.

Method 2: The Interpretability of Total and Subscales Scores

When item response data are found to be multidimensional, the interpretability of the total score as reflecting variation on a single construct is called into question. This is a valid but possibly unnecessary concern. Even in the presence of multidimensionality, total scale scores justifiably can be interpreted as demonstrated in Gustafsson and Aberg-Bengtsson (2010) and Reise, Moore, and Haviland (2010). Moreover, findings of multidimensionality do not guarantee that subscales can provide meaningful and reliable information about subdomains that is unique from the general construct.

In this section, we consider the question of TAS–20 total and subscale score interpretability in the presence of multidimensionality. By interpretable, we mean the degree to which total scores reflect a single construct versus being confounded by multidimensionality and the degree to which subscale scores

Table 2.—Application of the Haberman (2008) procedure to the Yates data set and four other published data sets examining the Twenty-Item Toronto Alexithymia Scale (TAS–20).

	Yates	Zhu S	Zhu C	Culhane A	Culhane H
SD TAS–20	11.5	9.2	11.0	10.5	12.2
SD DIF	5.9	4.7	5.7	5.5	6.3
SD DDF	4.3	3.5	3.6	4.6	5.1
SD EOT	4.4	3.9	4.3	4.4	4.7
r DIF–DDF	.63	.60	.74	.60	.59
r DIF–EOT	.27	.25	.31	.02	.22
r DDF–EOT	.35	.26	.37	.17	.21
αTAS–20	.85	.79	.84	.80	.84
αDIF	.86	.77	.82	.80	.83
αDDF	.77	.65	.62	.79	.82
αEOT	.60	.52	.66	.62	.60
H DIF	.71	.69	.76	.60	.70
H DDF	.72	.68	.86	.68	.65
H EOT	.42	.38	.44	.19	.31

Note. Zhu S = Zhu et al. (2007) student sample ($N = 870$); Zhu C = Zhu et al. (2007) clinical sample ($N = 179$); Culhane A = Culhane et al. (2009) Anglo student sample ($N = 367$); Culhane H = Culhane et al. (2009) Hispanic student sample ($N = 241$); *SD* = standard deviation; *r* = correlation between subscales; α = coefficient alpha; H = Haberman's proportional reduction in mean squared error (i.e., reliability) based on total scores rather than subscales; DIF = difficulty identifying feelings; DDF = difficulty describing feelings; EOT = externally oriented thinking.

reflect a construct that is unique from the construct represented by the total score. To accomplish this, we use model-based reliability estimation (Brunner & Süß, 2005; Miller, 1995; Raykov, 1997) from CFA. The specific latent variable model used here is a bifactor model (Holzinger & Swineford, 1937)—a multidimensional structural model specifying that each item on a measure is an indicator of a single factor (labeled the "target" dimension), and each item also is an indicator of one (or more) orthogonal group factors. The group factors in a bifactor model (in this example, DIF, DDF, EOT) represent common sources of variance among the items, controlling for the common variance explained by the general factor, alexithymia.

Specifically, to judge total score interpretability, an index is needed to estimate the percentage of variance in observed scores due to variance on a single common latent variable (i.e., the target construct). One approach to accomplishing this is to estimate a bifactor structure and compute indexes such as omega and omega hierarchical (omegaH; McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005). We are by no means the first to make these arguments (Revelle & Zinbarg, 2009). In an evaluation of the TAS–20, in fact, Gignac et al. (2007) pointed out exactly the same thing, but they cited a different source (i.e., Hancock & Mueller, 2001). In the context of other psychological measures, similar claims are made in Reise et al. (2010), Gustafsson and Aberg-Bengtsson (2010), Brown, Finney, and France (2011), and Zinbarg, Barlow, and Brown (1997).

There are several ways to write an equation for omega and omegaH depending on whether one is working with an exploratory or confirmatory bifactor model, and the precise interpretation depends on whether the factor pattern is based on the analysis of covariances, Pearson correlations, or tetrachoric/polychoric correlations. Nevertheless, the basic idea is to estimate the proportion of total score variance that can be attributed to all common factors (omega) and the proportion of total score variance that can be attributed to a single common factor (omegaH). If alexithymia, for example, were the general factor in a bifactor pattern (loadings standardized), and DIF, DDF, and EOT items loaded on three group factors, respectively, we could write:

$$\text{omega} = \frac{\left(\sum_{i=1}^{20}\lambda_{\text{iAlex}}\right)^2 + \left(\sum_{i=1}^{7}\lambda_{\text{iDIF}}\right)^2 + \left(\sum_{i=8}^{12}\lambda_{\text{iDDF}}\right)^2 + \left(\sum_{i=13}^{20}\lambda_{\text{iEOT}}\right)^2}{\left(\sum_{i=1}^{20}\lambda_{\text{iAlex}}\right)^2 + \left(\sum_{i=1}^{7}\lambda_{\text{iDIF}}\right)^2 + \left(\sum_{i=8}^{12}\lambda_{\text{iDDF}}\right)^2 + \left(\sum_{i=13}^{20}\lambda_{\text{iEOT}}\right)^2 + \sum_{i=1}^{20}\left(1-h_i^2\right)} \quad (3)$$

$$\text{omegaH} = \frac{\left(\sum_{i=1}^{20}\lambda_{\text{iAlex}}\right)^2}{\left(\sum_{i=1}^{20}\lambda_{\text{iAlex}}\right)^2 + \left(\sum_{i=1}^{7}\lambda_{\text{iDIF}}\right)^2 + \left(\sum_{i=8}^{12}\lambda_{\text{iDDF}}\right)^2 + \left(\sum_{i=13}^{20}\lambda_{\text{iEOT}}\right)^2 + \sum_{i=1}^{20}\left(1-h_i^2\right)} \quad (4)$$

where $(1 - h^2)$ is an item's unique variance, and the subscripts refer to item order in Table 1. The omega coefficient in Equation 3 is an estimate of the percentage of variance in observed total scores due to all sources of common variance. It is analogous to coefficient alpha in that its value is affected by all sources of common variance; however, the interpretation of the model-based omega coefficient as a reliability estimate does not depend on the assumption of tau equivalence (see Graham, 2006).

Most important, the omegaH in Equation 4 has several interpretations, but the one we are most concerned with here is that it is an estimator of the percentage of test score variance accounted for by variation on the general factor. OmegaH tells one directly the degree to which total scores are interpretable as indicators of the target construct of interest. To the degree that its value is high, relative to the percentage of variance explained by the group factors, one can proceed confidently with interpretation of the total raw score as reflecting the intended target construct. Moreover, comparison of omega to omegaH is useful in revealing the degree to which an estimate of reliability is inflated due to multidimensionality. In short, omega estimates the reliability of the multidimensional composite total score, whereas omegaH indicates the degree to which the total scores reflect variation on the target dimension. It is the latter that is most useful in judging the degree to which scores reflect a single latent variable.

We now demonstrate application of confirmatory bifactor modeling and computation of omega and omegaH with the Yates data. To contextualize these analyses, first we display a confirmatory three-correlated-factors model because of its considerable support and general acceptance. In the first column of Table 3 are the results of a CFA using maximum likelihood estimation based on a polychoric correlation matrix using EQS 6.1 (Bentler, 2006). Robust model fit indexes were Satorra and Bentler scaled $\chi^2(167) = 1{,}528$, $p < .001$, comparative fit index (CFI) = .939, standardized root mean sqaure residual (SRMR) = .084, and root mean square error of approximation (RMSEA) = .071 (.068–.074). By simply inspecting the loading pattern in the confirmatory model, its attractiveness as a conceptual frame is clear, but the fit is borderline if one strictly adheres to conventional benchmarks. Such benchmarks, however, have never been shown to be applicable for the evaluation of bifactor models based on polytomous items (see West, Taylor, & Wu, 2012).

In the second column of Table 3 are the loadings from a confirmatory bifactor model specifying one general and three group factors using EQS 6.1 (Bentler, 2006). The robust fit of this model was Satorra–Bentler $\chi^2(150) = 1{,}050$, $p < .001$, CFI = .960, SRMR = .057, and RMSEA = .061 (.058–.065). These values all are improvements over the confirmatory three-correlated-factors model. This is unsurprising given that the correlated-factors model is nested within the bifactor (Rindskopf & Rose, 1988).

In Row A are the sum of loadings down each column, then squared. In Row B is the sum of the error variances (uniqueness) for each item. Adding Rows A and B produces an estimate of the modeled variance $(115.2)^2$ that completes the denominator of Equations 3 and 4. Coefficient omega simply is the sum of row A (104.1) divided by 115.2, which equals .90. Dividing the sum-of-loadings squared (Row A) value for the general factor (88.4) by the estimated variance (115.2) yields an omegaH

[2]Note that within rounding, this value is equal to the sum of the tetrachoric correlation matrix.

TABLE 3.—Confirmatory models of the Twenty-Item Toronto Alexithymia Scale: Three correlated factors and bifactor.

		Three Factors				Bifactor			
Item	Domain	DIF	DDF	EOT		Alex	DIF	DDF	EOT
1	DIF	.76				.78	.02		
3	DIF	.62				.54	.44		
6	DIF	.76				.73	.21		
7	DIF	.68				.58	.80		
9	DIF	.78				.76	.17		
13	DIF	.85				.84	.10		
14	DIF	.74				.73	.07		
2	DDF		.80			.74		.32	
4	DDF		.69			.55		.53	
11	DDF		.72			.67		.22	
12	DDF		.61			.51		.34	
17	DDF		.53			.43		.36	
5	EOT			.37		.09			.37
8	EOT			.41		.34			.29
10	EOT			.63		.14			.59
15	EOT			.35		.27			.24
16	EOT			.21		.13			.13
18	EOT			.53		.09			.54
19	EOT			.74		.14			.80
20	EOT			.20		.35			.09
		1	.83	.26	A	88.4	3.28	3.13	9.30
		.83	1	.39	*B*	11.1			
		.26	.39	1	C	115.2			
					OmegaH	.76	.03	.03	.08

Note. Part A is the sum of the factor loadings for each factor, then squared. Part B is the sum of the error variances, and Part C is the sum of Parts A and B; it is the denominator for omega and omegaH. OmegaH is Part A divided by Part C. DIF = difficulty identifying feelings; DDF = difficulty describing feelings; and EOT = externally oriented thinking.

estimate of .76. Substituting in the remaining Row A values one at a time for the group factors in the numerator shows that .03%, .03%, and .08% of the remaining modeled total score variance on the TAS–20 is attributable to variation on the group factors (error is 10%). An omegaH estimate of .76 for the general factor indicates that .76/.90 = 84% of the reliable variance in TAS–20 scores is due to the general factor, suggesting that alexithymia is the only meaningful influence on true score variation. These findings are consistent with the scale author's clear and persistent recommendation to score the TAS–20 as a univocal measure.

Earlier, we used omega and omegaH to determine the degree to which interpretation of the total score was muddled by multidimensionality (i.e., confounded by subdomain constructs). Rarely is it discussed, however, that the results from the bifactor model can be used to address the opposite question; namely, to what degree is interpretation of subscale scores confounded by the general factor? In fact, we can use the bifactor model results to estimate both the reliability of subscale scores and the degree to which subscales provide reliable information that is unique from the general factor. The first step is to estimate omega for each subscale separately. For example, omega for the DIF subscale would be computed as:

$$\text{omega} = \frac{\left(\sum_{i=1}^{7}\lambda_{\text{iAlex}}\right)^2 + \left(\sum_{i=1}^{7}\lambda_{\text{iDIF}}\right)^2}{\left(\sum_{i=1}^{7}\lambda_{\text{iAlex}}\right)^2 + \left(\sum_{i=1}^{7}\lambda_{\text{iDIF}}\right)^2 + \sum_{i=1}^{7}\left(1-h_i^2\right)} \quad (5)$$

Notice that only the loadings and error terms for the first seven items are relevant to this calculation. The omega estimates for DIF, DDF, and EOT subscales were .92, .82, and .66, respectively. The second step is to calculate the unique reliability for each subscale after controlling for the general factor. This is accomplished by removing the first term from the numerator in Equation 5, as shown in Equation 6 for the DIF subscale.

$$\text{omegaS} = \frac{\left(\sum_{i=1}^{7}\lambda_{\text{iDIF}}\right)^2}{\left(\sum_{i=1}^{7}\lambda_{\text{iAlex}}\right)^2 + \left(\sum_{i=1}^{7}\lambda_{\text{iDIF}}\right)^2 + \sum_{i=1}^{7}\left(1-h_i^2\right)} \quad (6)$$

Note that we now use the term omegaS to denote the estimate of reliability for a subscale after controlling for the general factor. In these data, these values are .11, .22, and .53 for DIF, DDF, and EOT, respectively. These omegaS values indicate that DIF and DDF contain little reliable variance that is unique from the general factor. Simply stated, the overwhelming majority of the reliable variance on DIF and DDF subscales can be attributable to the general factor (alexithymia). This is not the case for the EOT subscale, however, considering that its omega was .66, and, thus, the majority of reliable variance on EOT subscale scores (.53/.66 = .80) is independent of the general factor.

METHOD 3: MODELING MULTIDIMENSIONAL DATA IN SEM

As argued by numerous scholars, SEM is the appropriate analytic tool when researchers wish to evaluate the disattenuated (i.e., controlling for measurement error) interrelations among a network of constructs. As with all latent variable modeling, SEM requires the latent variable to reflect variation on a single construct; that is, to reflect the common variance among the indicators. Stated differently, SEM assumes that the latent variable indicators reflect a single common latent variable or that all multidimensionality (e.g., correlated errors) has been specified in the model by allowing correlated residuals or specifying more complicated multidimensional measurement models (e.g., a second-order or bifactor model; Little, Cunningham, Shahar, & Widaman, 2002). This is a strict assumption, and as noted, few psychological measures will yield strictly unidimensional data.

Violating unidimensionality can have severe consequences in SEM; treating multidimensional data as if it were unidimensional, for example, is a form of model misspecification that can lead to biased parameter estimates (i.e., loadings too high, error estimates too low), and more important, biased parameter estimates lead to biased estimates of the relationships among latent variables. Such bias can defeat the primary advantage of SEM; namely, to accurately gauge the true relations among constructs represented as latent variables. Thus, if a measure generates multidimensional item response data, a researcher needs to consider whether it still is appropriate to specify a unidimensional measurement model or whether one needs to consider alternative, and very much more complicated, potentially unreplicable, multidimensional latent variable model specifications (e.g., second-order or bifactor measurement models; see Chen et al., 2006).

In the previous sections, we have shown that for TAS–20 item responses (a) DIF and DDF items load very highly on a

TABLE 4.—Confirmatory bifactor model parameter estimates of the Twenty-Item Toronto Alexithymia Scale based on Yates data (four group factors) and Gignac Model 5B (five group factors).

#[a]	Yates 4 Group Factors					#[a]	Gignac et al. (2007) Model 5B					
1	.79	.02				1	.51	.48				
3	.51	.60				3	.29	.36				
6	.73	.20				6	.60	.26				
7	.60	.60				7	.45	.43				
9	.76	.18				9	.48	.57				
13	.84	.13				13	.65	.40				
14	.73	.09				14	.54	.29				
2	.74		.34			2	.46		.56			
17	.42		.33			17	.50		.42			
11	.68		.20			11	.59		.24			
12	.51		.30			12	.62		.25			
4	.54		.56			4	.43		.71			.51
10	.15			.60		10	.51			.34		
18	.10			.56		18	.36			.43		
19	.15			.78		19	.47			.73		.17
15	.28			.24		15	.10				.23	
16	.14			.11		16	.09				.45	
5	.09				.46	5	.25				.16	.28
8	.34				.42	8	.39				.27	
20	.26				.11	20	.29				.57	

[a]Item order changed to be consistent with Gignac et al. (2007).

general factor and only modestly on group factors in a bifactor model, and (b) the general factor in a bifactor model accounts for 84% of the reliable TAS–20 score variation. Taken together, these findings generally support TAS–20 researchers modeling alexithymia with a unidimensional measurement model in SEM research. Nevertheless, to empirically evaluate this assertion, we conduct two very simple analyses; in each, taking a correlation matrix generated from a plausible, well-fitting multidimensional model, and then forcing the data into a unidimensional measurement model. Our expectation is that, in the case of the TAS–20, forcing multidimensional data into a unidimensional measurement structure biases validity coefficient estimates only minimally (e.g., relative bias less than 10%).

Table 4 shows two highly multidimensional CFA solutions. The solution on the left is based on the Yates data. In this model, there is a single general factor reflecting alexithymia and four group factors because the EOT factor has been divided into two factors as suggested in the literature (Meganck et al., 2008; Müller et al., 2003). Robust fit statistics are CFI = .953, SRMR = .067, and RMSEA = .066 (.062–.069). The second structure has even more dimensions. This is a confirmatory structure published in Gignac et al. (2007) where in addition to breaking EOT into two factors, a fifth group factor is specified to account for additional method variance. The (nonrobust) fit for this model is CFI = .947, SRMR = .043, RMSEA = .046 (.037–.056). Note also that the Gignac et al. solution is based on Pearson correlations and the Yates data are based on polychoric correlations. That might be one reason why loadings generally are higher in the Yates model. Note, too, that the models break apart the EOT factor in two different ways.

These confirmed multidimensional models serve as the basis for generating correlation matrices that, in turn, were forced into a unidimensional SEM measurement model (i.e., all 20 items specified as indicating a single latent variable). The specific procedure for each multidimensional data structure was as follows.

1. We specified an additional latent variable with three indicators (Items 21, 22, and 23) with standardized loadings of .70, .75, and .80 (to identify the latent variable). This latent variable has a variance of 1.0 and serves as a criterion variable.
2. For each model, we conducted analyses using four different values for the structural coefficient between the criterion latent variable and the general factor in the bifactor model. These values were 0, .2, .4, and .6. Note that these structural coefficient values sometimes are referred to as validity coefficients.
3. For each model and true validity coefficient combination, the complete factor loading matrix (bifactor loadings plus the criterion variable loadings) was transformed into a correlation matrix using the simple relation:

$$R = \lambda \phi \lambda^T \tag{7}$$

In Equation 7, λ is the matrix of factor loadings (23 by 6 for the Yates model, 23 by 7 for the Gignac et al. 2007 model), and ϕ is a matrix of factor intercorrelations. The resulting reproduced correlation matrix (R) then is treated as a true population correlation matrix generated from a known (bifactor) multidimensional structure and criterion variable.

4. For each reproduced correlation matrix from Equation 7, we then estimated a unidimensional model using EQS 6.1 (Bentler, 2006) by specifying that each of the TAS–20 items reflects a single latent variable, the criterion latent variable has three indicators, and the correlation between alexithymia and criterion latent variable is freely estimated. It is this model that is critical in judging the effects of forcing multidimensional data into a unidimensional model. To the degree that multidimensionality affects the factor loading estimates (making them different from their values on the general factor in the bifactor model), we should see bias in the structural coefficients. To the degree that there is no or limited bias in the validity coefficients, multidimensionality has no demonstrable practical consequence, at least in terms of the present structural representation.

Results for the Yates model were slightly better than for the Gignac et al. (2007) model. Specifically, in the Yates model, validity coefficients were .000, .195, .390, and .585, when the true values were 0, .2, .4, and .6, respectively. In the Gignac et al. model, validity coefficients were .000, .182, .364, and .546, when the true values were 0, .2, .4, and .6, respectively. Clearly, there is some degree of absolute bias here, especially as the true validity gets larger. The relative bias, however, is never larger than 10%; thus, it is reasonable to conclude that in terms of estimating a single structural (validity) coefficient, the degree of absolute bias caused by the model misspecification is relatively low and likely of no substantive consequence. In short, fitting the misspecified unidimensional model by treating the TAS–20 as a single latent variable has few practical implications in terms of validity coefficient estimates in SEM. This remains true even though the unidimensional model has been shown to display "unacceptable" fit to TAS–20 data across multiple studies.

DISCUSSION

When item response data are strictly unidimensional (a) total scores can be unambiguously interpreted as reflecting variation

on a single common latent variable, (b) the creation of subscales is psychometrically indefensible (see Bollen & Lennox, 1991), and (c) specifying a unidimensional measurement model in SEM is appropriate (no parameter bias due to misspecification). At best, however, the assumption that a single common latent variable explains item response data (unidimensionality) is a convenient fiction, sometimes useful in applied contexts but often not. Unfortunately, the mere "confirmation" of a multidimensional structure is a necessary first step for determining whether a researcher needs to account for this multidimensionality (i.e., through scoring subscales and representing constructs in SEM) or ignore it. This "confirmation," however, does not provide sufficient decision-making guidance.

In this report, we considered three psychometric methods to address four distinct questions, all tied together under the common theme of exploring the consequences of multidimensionality in terms of scale scoring and specifying a measurement model in SEM. These approaches can be implemented easily and are applicable to a wide range of psychological measures. The techniques presented here, for example, are relevant to any measure where researchers debate its multidimensional structure and ask whether total or subscale scores should be reported or used in research or clinical practice. They are applicable, too, to measures that were developed to assess a single target construct (in the case of our running example, "alexithymia") but include multiple domains of item content to better represent the construct. For such measures, findings of multidimensionality are inevitable. Next, we review each of the three demonstrations and also suggest contexts in which each of these methods is most important, and provide guidance for applied researchers on how best to interpret their results.

Method 1: The Haberman (2008) Procedure

The Haberman (2008) procedure has its foundation in large-scale educational testing and is based on classical test theory principles. As a consequence, the procedure is applicable across a broad range of measures where researchers question the value of scoring and reporting both total and subscale scores. As we noted previously, under certain conditions (e.g., high correlations between subscales and total scores, low reliability for subscales), it is possible that the total score provides a relatively more precise indicator of subscale true scores than the subscale score, and, thus, the subscale score provides no "added value" and should not be reported or interpreted. The Haberman procedure allows researchers to empirically determine when this is the case. In the present data, all three proposed TAS–20 subscales, DIF, DDF, and EOT, cleared this relatively low hurdle, and, thus, there is some support, albeit modest for DIF and DDF, for scoring the subscales and using them in research.

In our example analysis, we have promoted the use of the Haberman procedure to make a simple (yes–no) decision on whether subscale scores should be reported (or by extension, used in research or policy and decision-making). But what should a researcher do when a subscale clears this hurdle, but the ratio of $PRMSE_{TOT}$ to $PRMSE_S$ is high (e.g., $> .80$), as in this case, suggesting that subscale scores do not provide much additional precision that could not be garnered from knowing an individual's total score? Our answer depends on context. For educational assessment researchers, we note that the ratio of $PRMSE_{TOT}$ to $PRMSE_S$ can be used to inform decisions regarding the augmenting of subscale scores—a method of borrowing strength from the total score so that the subscale is more reliable (Sinharay, Haberman, & Wainer, 2011). In psychological assessment more generally, where to our knowledge the concept of score augmentation has never been considered, we suggest not attempting to interpret the value of the $PRMSE_{TOT}$ to $PRMSE_S$ ratio.

Our reasoning is that the $PRMSE_{TOT}$ to $PRMSE_S$ ratio simply is a ratio of two estimates of how precisely individual differences can be assessed on a subdomain—one based on subscale scores and the other on total scores. This ratio, although important, does not address the critical question of whether total or subscale scores are influenced by multiple sources of variance. To address those issues, we suggest the latent variable modeling-based techniques discussed next. Nevertheless, in cases where latent variable modeling is not possible, perhaps because a measure has no clear latent structure, we suggest that Equation 2, which yields an estimate of the correlation between true total scores and true subscale scores, is more directly informative in terms of judging the distinctiveness of a subscale score. In this example, these values were .84, .85, and .49 for DIF, DDF, and EOT, respectively. These values suggest that individual differences in either DIF or DDF are highly redundant with individual differences on alexithymia, but individual differences on EOT are only moderately redundant with individual differences on alexithymia.

Finally, we note that the "added value" of subscales issue addressed by the Haberman (2008) procedure, perhaps, is an issue most pertinent to large-scale educational testing, where instruments are designed, analyzed, and scored by professional psychometricians and where such scores routinely are used to make critical individual, school-level, and societal decisions of great consequence. Sinharay, Puhan, and Haberman (2011) presented differences between achievement tests and personality and clinical scales, noting, in particular, that personality scales are intended to measure narrower constructs. They suggested that the problems with unreliable subscales and high correlations between subscales and total test scores might not be so prevalent outside of educational testing.

We appreciate their arguments, but also believe that there is a superabundance of examples of personality and psychopathology measures where an essentially unidimensional domain has been broken up unnecessarily into subscales. For example, Reid, Garos, and Carpenter (2011) reported on a CFA of their Hypersexual Behavior Inventory, in which they confirmed three correlated factors, which when scored as subscales, have subscale-to-total score correlations ranging from $r = .92$ to .95. This example of new scale development is a prime case in which the Haberman procedure can serve psychological assessment more generally as a minimal hurdle to clear when considering the development or reporting of subscale scores, even when multidimensionality is confirmed.

Method 2: Model-Based Reliability and OmegaH and OmegaS

Demonstrating that a subscale provides added value and provides meaningful information are two very different conclusions, but justification must precede interpretation. Given that a subscale, at least, has cleared the Haberman test hurdle, it is important to explore two more questions: (a) to what extent the

total score is interpretable as a measure of a single common construct, and (b) to what degree subscales provide reliable measures of constructs independently of the general construct. To address these questions, we used model-based reliability estimates computed on the factor loadings from a confirmatory bifactor model.

As noted, multiple authors have commented on the fact that even when data are highly multidimensional, scale scores still can predominantly reflect a single latent variable (e.g., Gustafsson & Aberg-Bengtsson, 2010). To determine the degree to which this is true for the TAS–20, we estimated omegaH based on a confirmatory bifactor model[3] with one general and three group factors. This index is an estimate of the variance in total scores that can be attributable to the general factor running across the items. In these data, omegaH was estimated to be 76%. Further analyses revealed that the DIF, DDF, and EOT latent variables accounted for 3%, 3%, and 8% of variance, respectively, in TAS–20 scores, and error was 10%. Thus, the reliable variation in TAS–20 scores is almost entirely attributable to a single, general latent variable, ostensibly reflecting alexithymia. We see no empirical reason why researchers could not be confident in interpreting TAS–20 scores as reflecting variation on a single construct.

Model-based reliability estimation based on omegaS also was used to determine the degree to which proposed subscale scores are reliable after controlling for the reliable variance due to the general factor (alexithymia). Unsurprisingly, given the high correlation between DIF and DDF, and their demonstrated redundancy with total scale scores, the reliability of these subscale scores is almost entirely attributable to systematic variation on the general factor and not to systematic variation on the unique subscale constructs. Clearly, to meaningfully interpret correlates of these subscales, a researcher must use bifactor models to separate the general construct from the specific and SEM models to simultaneously study the unique correlates of both (see Chen et al., 2006, for more details). On the other hand, the EOT subscale is not highly correlated with either DIF or DDF subscale scores or, in fact, to the total TAS–20 score. Thus, it also is not surprising that although the reliability of EOT scores is relatively low, at least, most of that reliability is attributable to systematic variation on a construct unique from alexithymia.

The limitations of coefficient alpha, and the virtues of model-based reliability estimation, as well as the utility of omegaH are well documented in the psychometric literature. Nevertheless, in personality and clinical assessment, coefficient alpha remains the only index that is universally reported, whereas alternatives such as omega, omegaH, or omegaS rarely are reported. The problem, however, is not a lack of modeling research in these domains. Indeed, concerns about the multidimensionality of popular personality and clinical instruments and arguments about scoring the total score versus subscales, typically, arise from multiple CFA studies published in personality assessment journals that demonstrate a multifactor model fits better than a unidimensional model, as is the case for the TAS–20.

We speculate that the reason why indexes such as omegaH and omegaS seldom are reported (and in turn, why questions about the interpretability of total and subscale scores are not addressed routinely) has more to do with the type of CFA typically employed, more so than too few modeling studies. Specifically, most CFA studies either evaluate a correlated-factors or a second-order model. In the former, there is no statistical representation of a common trait running through the items, and correlations (i.e., redundancy) among factors are hidden in the factor correlation matrix. In the latter, a common factor is represented by a second-order latent variable, but it is proposed to explain correlations among the primary factors, not among items. Moreover, in a second-order model, there are no direct effects between the second-order factor and the items—the only effect is indirect through the primary factors.

It is only the bifactor representation that stipulates there is a common factor running through all the items and allows researchers to partition the common variance into that due to the general factor and that due to additional common group factors. In turn, it is the bifactor model that provides the framework for the computation and interpretation of omegaH and omegaS. But bifactor structural representations, either confirmatory or exploratory, seldom have been employed. Thus, the requirement of a bifactor model to compute indexes such as omegaH and omegaS might appear to be a limitation, but it should not be. It must be recognized that both the correlated-factors and second-order models are nested within a bifactor, as noted earlier. Thus, whenever a researcher conducts a CFA and finds an acceptable fit for a correlated-factors or a second-order model, it is highly likely, if not certain, that a bifactor representation provides a superior fit. As such, we argue that whenever a CFA is proposed as evidence of a multidimensional structure with correlated factors, a bifactor model also should be reported with corresponding omegaH and omegaS values. In this way, researchers will be in a better position to judge the degree to which multidimensionality affects interpretation of both total and subscale scores.

That said, we have no specific advice in terms of benchmarks for evaluating omegaH or omegaS. Tentatively, we can propose that a minimum would be greater than .50, and values closer to .75 would be much preferred, but that is a subjective guideline. In fairness, we must also point out some limitations of omegaH and omegaS. First, as model-based statistics, their value is dependent on the model estimated. If in the present case, more or fewer group factors were specified, somewhat different values of omega values would result. Second, omegaH tends to increase with test length (see Gustafsson & Aberg-Bengtsson, 2010) and is influenced by the specific structure of multidimensionality (Reise, Scheines, Widaman, & Haviland, 2012). For example, all else being equal (e.g., test length), omegaH will be higher in a test with many small group factors as opposed to one with a few large group factors. Third, as noted in several sources (e.g., Reise et al., 2010), items that cross-load on more than one group factor can seriously distort all factor loading estimates in bifactor models, especially confirmatory models. In turn, if the loadings are poorly estimated, then any estimate of omegaH or omegaS must be treated with skepticism.

Method 3: Specifying a Measurement Model in SEM

When researchers primarily are interested in estimating the relations among constructs, SEM is an attractive data analytic

[3] A confirmatory model is not a requirement for computation of either coefficient (see McDonald, 1999); exploratory bifactor methods such as the Schmid–Leiman (Schmid & Leiman, 1957) target bifactor rotation (Reise, Moore, & Maydeu-Olivares, 2011) and the Jennrich–Bentler (Jennrich & Bentler, 2011) bifactor rotation could have been used as well (see Reise, in press).

approach. Although SEM requires no "scoring" of item responses, the total score versus subscale issue remains in the form of debates regarding how best to specify a measurement model. Would a parsimonious unidimensional model properly represent the construct, or must researchers specify a more complex structure such as a bifactor or second-order measurement model?

An emerging practice in SEM research is the specification of more complicated measurement models (e.g., bifactor) that better take into account the multidimensional nature of item responses derived from popular measures (e.g., see Ackerman, Donnellan, & Robins, 2012; Brown et al., 2011; Chen et al., 2006; Chen et al., 2012; Simms, Gros, Watson, & O'Hara, 2008; Thomas, 2012). Although such models inevitably provide a better fit than unidimensional or correlated factors representations, this does not automatically imply that the added complexity of such models results in any practical gains in precision. This is especially true when item response data, although multidimensional, are characterized by a strong general factor, such as the TAS–20 (see Reise et al., 2012).

For this reason, and with clinical diagnoses in mind as a criterion, Thomas (2012) argued, "Simulation work is needed to compare the diagnostic benefits between the simple structure and bifactor models" (p. 110). We agree that more complicated models need to be justified not merely on their better fit but also on their demonstrated applied advantages. Accordingly, in this report, we presented a fairly simple method of evaluating the consequences of forcing data with a multidimensional latent structure into a unidimensional measurement model.

Technically, our approach is not a simulation because we have no need to explore the effects of sampling error on our parameter estimates. Rather, our approach is based on "comparison modeling" (Reise, Cook, & Moore, in press; Reise, Morizot, & Hays, 2007). In comparison modeling, the parameter estimates derived from a well-fitting more complicated model (e.g., bifactor) are considered a "true" representation of the population structure. Then, we simply convert the "true" parameters into a correlation matrix, fit a more restricted model (e.g., unidimensional), and observe the effects on important parameter estimates.

In this article, comparison modeling was used to judge the bias in validity coefficients that occurs when confirmed multidimensional (bifactor) structures are forced into a unidimensional model. This technique easily can be adapted to study the effects of placing restrictions on any published factor structure and should be used whenever a researcher is debating how best to use item responses to represent a construct in SEM. Of course, it must be recognized that reported confirmatory models, as used here, are "cleaned up" representations (e.g., small cross-loadings are forced to be zero in confirmatory modeling).

For the TAS–20 data, we concluded that although item responses, clearly, are better fit by multidimensional representations, there is no great loss in the accuracy of validity coefficients as a function of treating the data as unidimensional. This is compelling evidence that the item responses are "essentially unidimensional." In the case of the TAS–20, this a relief—sorting through several competing "confirmed" TAS–20 structures to inform the correct measurement model for our research, and then hoping that the data "fit" a highly restricted multidimensional model such as the bifactor, where no cross-loadings are allowed, seems a tedious exercise for a measure with such a strong general factor, and subscales that provide little if any reliable information that is unique from the general factor.

On the other hand, we recognize that even when unidimensional measurement models might be good enough for most practical purposes, there are research contexts where more complex measurement models are necessary; for example, if a TAS–20 researcher wished to test a theory regarding the unique contribution of general (alexithymia) and group factors (DIF, DDF, and EOT) to the prediction of important criteria. Moreover, likely there are many psychological measures where the type of analyses suggested here will reveal that forcing a unidimensional measurement model onto multidimensional data, indeed, would bias important parameter estimates in a serious way. In such cases, we fully support the use of complex representations; however, mere "better fit" is not a sufficient justification—the more complex model must really make a difference.

Summary

We presented three methods for evaluating the applied effects of multidimensionality for personality and clinical measures. The Haberman procedure is used to determine if a subscale score provides a more accurate index of subscale true scores relative to the total test score. To the degree that subscale scores are reliable and less correlated with total scores, subscales will pass the Haberman test. When a subscale fails the Haberman test, decision making is entirely clear—do not calculate, report, or use subscale scores.

We next demonstrated model-based reliability coefficients omegaH and omegaS, which depend on the estimation of a bifactor model. These coefficients are used to address the questions of to what degree total scores are interpretable as an indicator of a single construct, and to what degree subscale scores reflect a construct that is independent of the general construct measured by an instrument. OmegaH will be high as a joint function of the items' relative loadings on the general to the group factor and to the number of items. OmegaS will be high to the degree that items load higher on their group factor relative to the general factor, or, stated differently, to the degree that subscale scores are uncorrelated with total scores. Unfortunately, we cannot offer definitive interpretive ("acceptability") guidelines for either omegaH or omegaS. We, however, can say that if the primary goal of the instrument is to measure a single individual construct, one would want omegaH to be as high as possible. We also can state that subscale scores are more useful for research purposes to the degree that omegaS is high (or the subscale score is not highly correlated with the total score). There are limits, however; if omegaS, for example, equaled 1.0 or a subscale had a correlation of zero with the total score, then it would be impossible to claim that it is a subscale representing a more narrow aspect of a broader construct.

The third method was comparison modeling, where we illustrated the use of SEM to judge the effect of forcing multidimensional data structures into a unidimensional measurement model on the bias in structural coefficients. We, again, cannot provide clear benchmarks for "good" versus "bad." In this study, we used a criterion of 10% bias as acceptable. Some researchers might tolerate higher levels, but still others will insist that no level of bias is acceptable. For this latter group, nothing but a well-fitting multidimensional measurement model will suffice. Regardless of one's "good–bad" standard, the comparison

modeling procedure provides a way for researchers who want to directly evaluate the effects of multidimensionality on important parameters, such as factor loadings or structural coefficients. The only requirement is that the researcher has a well-fitting multidimensional model to serve as the comparison.

Finally, our TAS–20 analyses were largely for demonstration purposes. The substantive findings, however, might be useful to researchers as they evaluate published TAS–20 CFAs and studies in which the authors have drawn conclusions about the relationship of alexithymia to an external correlate. They could be useful, too, as researchers plan new TAS–20 structural evaluations and SEM studies, as well as in the evaluation of observer alexithymia measures as demonstrated by Reise et al. (2010).

Acknowledgments

Portions of this research were made possible by a predoctoral advanced quantitative methodology training grant (#R305B080016) awarded to UCLA by the Institute of Education Sciences of the U.S. Department of Education. The authors thank Peter Bentler for reviewing the factor analyses in this study and Rob Meijer and Michael Furr for their comments on earlier drafts.

References

Ackerman, R. A., Donnellan, M. B., & Robins, R. W. (2012). An item response theory analysis of the Narcissistic Personality Inventory. *Journal of Personality Assessment*, *94*, 141–155.

Al-Turkait, F. A., & Ohaeri, J. U. (2010). Dimensional and hierarchical models of depression using the Beck Depression Inventory–II in an Arab college student sample. *BMC Psychiatry*, *10*, 60.

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The Twenty-Item Toronto Alexithymia Scale I: Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*, 23–32.

Bagby, R. M., Taylor, G. J., Quilty, L. C., & Parker, J. D. A. (2007). Reexamining the factor structure of the 20-Item Toronto Alexithymia Scale: Commentary on Gignac, Palmer, and Stough. *Journal of Personality Assessment*, *89*, 258–264.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory–II*. San Antonio, TX: Psychological Corporation.

Bentler, P. M. (2006). *EQS 6.1 structural equations program manual*. Encino, CA: Multivariate Software.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equations perspective. *Psychological Bulletin*, *110*, 305–314.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2012, July 16). On the factor structure of the Beck Depression Inventory–II: G is the key. *Psychological Assessment*. Advance online publication. doi:10.1037/a0029228

Brown, A. R., Finney, S. J., & France, M. K. (2011). Using the bifactor model to assess the dimensionality of the Hong Psychological Reactance Scale. *Educational and Psychological Measurement*, *71*, 170–185.

Brunner, M., & S̈ß, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement*, *65*, 227–240.

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

Culhane, S. E., Morera, O. F., Watson, P. J., & Millsap, R. E. (2009). Assessing measurement and predictive invariance of the Toronto Alexithymia Scale–20 in U.S. Anglo and U.S. Hispanic students. *Journal of Personality Assessment*, *91*, 387–395.

Gignac, G. E., Palmer, B. R., & Stough, C. (2007). A confirmatory factor analytic investigation of the TAS–20: Corroboration of a five-factor model and suggestions for improvement. *Journal of Personality Assessment*, *89*, 247–257.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*, 930–944.

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97–121). Washington, DC: American Psychological Association.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Joreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Haviland, M. G., & Reise, S. P. (1996a). A California Q-set alexithymia prototype and its relationship to ego-control and ego-resiliency. *Journal of Psychosomatic Research*, *41*, 597–608.

Haviland, M. G., & Reise, S. P. (1996b). Structure of the twenty-item Toronto Alexithymia Scale. *Journal of Personality Assessment*, *66*, 116–125.

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

Hund, A. R., & Espelage, D. L. (2005). Childhood sexual abuse, disordered eating, alexithymia, and general distress: A mediation model. *Journal of Counseling Psychology*, *52*, 559–573.

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

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173.

Lumley, M. A., Neely, L. C., & Burger, A. J. (2007). The assessment of alexithymia in medical settings: Implications for understanding and treating health problems. *Journal of Personality Assessment*, *89*, 230–246.

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

Meganck, R., Vanheule, S., & Desmet, M. (2008). Factorial validity and measurement invariance of the 20-item Toronto Alexithymia Scale in clinical and nonclinical samples. *Assessment*, *15*, 36–47.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, *2*, 255–273.

Müller, J., Bühner, M., & Ellgring, H. (2003). Is there a reliable factorial structure in the 20-item Toronto Alexithymia Scale? A comparison of factor models in clinical and normal adult samples. *Journal of Psychosomatic Research*, *55*, 561–558.

Osman, A., Barrios, F. X., Gutierrez, P. M., Williams, J. E., & Bailey, J. (2008). Psychometric properties of the Beck Depression Inventory–II in nonclinical adolescent samples. *Journal of Clinical Psychology*, *64*, 83–102.

Ouwens, M. A., van Strien, T., & van Leeuwe, J. F. J. (2009). Possible pathways between depression, emotional and external eating: A structural equation model. *Appetite*, *53*, 245–248.

Parker, J. D. A., Keefer, K. V., Taylor, G. J., & Bagby, R. M. (2008). Latent structure of the alexithymia construct: A taxometric investigation. *Psychological Assessment*, *20*, 385–396.

Parker, J. D. A., Taylor, G. J., & Bagby, R. M. (2003). The 20-Item Toronto Alexithymia Scale: III. Reliability and factorial validity in a community population. *Journal of Psychosomatic Research*, *55*, 269–275.

Quilty, L. C., Zhang, A. Z., & Bagby, R. M. (2010). The latent symptom structure of the Beck Depression Inventory–II in outpatients with major depression. *Psychological Assessment*, *22*, 603–608.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173–184.

Reid, R. C., Garos, S., & Carpenter, B. N. (2011). Reliability, validity, and psychometric development of the hypersexual behavior inventory in an outpatient sample of men. *Sexual Addiction and Compulsivity*, *18*, 30–51.

Reise, S. P. (in press). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*.

Reise, S. P., Cook, K. F., & Moore, T. M. (in press). A direct modeling approach for evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. R. Reise & D. Revicki (Eds.),

Handbook of item response theory and patient reported outcomes. London, England: Taylor & Francis.

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

Reise, S. P., Moore, T. M., & Maydeu-Olivares, A. (2011). Targeted bifactor rotations and assessing the impact of model violations on the parameters of unidimensional and bifactor models. *Journal of Educational and Psychological Measurement*, *71*, 684–711.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19–31.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). The effects of multidimensionality on predictive validity coefficients in structural equation modeling. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164412449831

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika*, *74*, 145–154.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*, 51–67.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Simms, L. J., Gros, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety*, *25*, E34–E46.

Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do adjusted subscales lack validity? Don't blame the messenger. *Educational and Psychological Measurement*, *71*, 789–797.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*, 29–40.

Taylor, G. J., Bagby, R. M., & Luminet, O. (2000). Assessment of alexithymia: Self-report and observer-rated measures. In R. Bar-on & J. D. A. Parker (Eds.), *Handbook of emotional intelligence: Theory, development, assessment, and application at home, school, and in the workplace* (pp. 301–319). San Francisco, CA: Jossey-Bass.

Taylor, G. J., Bagby, R. M., & Parker, J. D. A. (2003). The 20-Item Toronto Alexithymia Scale. IV. Reliability and factorial validity in different languages and cultures. *Journal of Psychosomatic Research*, *55*, 277–283.

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

Vanheule, S., Desmet, M., Groenvynck, H., Rosseel, Y., & Fontaine, J. (2008). The factor structure of the Beck Depression Inventory–II: An evaluation. *Assessment*, *15*, 177–187.

Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory–II. *Psychological Assessment*, *18*, 81–88.

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

Zhu, X., Yi, J., Yao, S., Ryder, A. G., Taylor, G. J., & Bagby, R. M. (2007). Cross-cultural validation of a Chinese translation of the 20-item Toronto Alexithymia Scale. *Comprehensive Psychiatry*, *48*, 489–496.

Zinbarg, R. E., Barlow, D. H., & Brown, T. A. (1997). Hierarchical structure and general factor saturation of the Anxiety Sensitivity Index: Evidence and implications. *Psychological Assessment*, *9*, 277–284.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's omega h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133.