JOURNAL OF PERSONALITY ASSESSMENT
https://doi.org/10.1080/00223891.2018.1449116

Routledge
Taylor & Francis Group

STATISTICAL DEVELOPMENT

Selecting the "Best" Factor Structure and Moving Measurement Validation Forward: An Illustration

Thomas A. Schmitt,[1,2] Daniel A. Sass,[2,3] Wayne Chappelle,[4] and William Thompson[2]

[1]Equastat; [2]NeuroStat Analytical Solutions; [3]Department of Management Science and Statistics, University of Texas at San Antonio; [4]U.S. Air Force School of Aerospace Medicine

ABSTRACT

Despite the broad literature base on factor analysis best practices, research seeking to evaluate a measure's psychometric properties frequently fails to consider or follow these recommendations. This leads to incorrect factor structures, numerous and often overly complex competing factor models and, perhaps most harmful, biased model results. Our goal is to demonstrate a practical and actionable process for factor analysis through (a) an overview of six statistical and psychometric issues and approaches to be aware of, investigate, and report when engaging in factor structure validation, along with a flowchart for recommended procedures to understand latent factor structures; (b) demonstrating these issues to provide a summary of the updated Posttraumatic Stress Disorder Checklist (PCL–5) factor models and a rationale for validation; and (c) conducting a comprehensive statistical and psychometric validation of the PCL–5 factor structure to demonstrate all the issues we described earlier. Considering previous research, the PCL–5 was evaluated using a sample of 1,403 U.S. Air Force remotely piloted aircraft operators with high levels of battlefield exposure. Previously proposed PCL–5 factor structures were not supported by the data, but instead a bifactor model is arguably more statistically appropriate.

ARTICLE HISTORY
Received 23 January 2018
Accepted 3 February 2018

In psychological and personality research, it is common to have an extensive research literature statistically evaluating a multitude of psychometric factor structures for theoretical and diagnostic purposes. Unfortunately, these factor exploration or confirmation endeavors have been and continue to be clouded by inconsistent and sometimes incorrect methods (e.g., relying too heavily on global measures of fit). These various factor models often lead to a published proliferation of excessively complex and contradicting factor structures that could be statistically unjustified and lack generalizability or predictive validity, resulting in insufficient clinical or practical relevance.

Although dimension reduction and model selection are essential for measure development, the central goal of modeling is description and prediction. Thus, it is important for researchers to be cognizant of balancing generalizability with accurately describing the underlying phenomenon (Cudeck & Henly, 1991; Preacher, Zhang, Kim, & Mels, 2013). This means explicitly stating and balancing scientific goals—searching for models that generalize (i.e., prediction and replicate) or models with verisimilitude (i.e., description and explanation)—and realizing that determining the true number of factors based on global model fit statistics (e.g., χ^2) is likely futile. In this context, researchers tend to rely too heavily on global model fit statistics, which often lead to arbitrary factor fishing, and overly complex and overfitted models (see Hayduk, 2014b). Users are frequently not cognizant that generalizability, goodness of fit, and overfitting are each a function of model complexity, and these "good fitting" models might fit the data well, but they can fail in explaining the data generating process and can exhibit poor predictive validity (see Myung & Pitt, 2002, Figure 11.4, p. 449; Pitt, Myung, & Zhang, 2002, Figure 2, p. 475; Preacher, 2006, Figure 3, p. 233).

Further, as model complexity increases, prediction error can drop to zero, but these models overfit the data and will typically have poor generalizability (Hastie, Tibshirani, & Friedman, 2009; James, Witten, Hastie, & Tibshirani, 2013). This phenomenon becomes more prevalent with smaller sample sizes and is known as the "bias-variance trade-off" (Hastie et al., 2009); with more complex models, bias decreases (approximation error decreases) and variance increases (prediction or estimation error increases). In practice, this means that complex models can fit well due to arbitrary properties of the model, but have very little to do with the optimal approximation to truth, and can display poor generalizability. This phenomenon is illustrated in Figure 1, where a point of maximum generalizability or maximum predictive validity is reached, after which models can exhibit excessive complexity and a decrease in generalizability due to overfitting the data. Thus, generalizability depicts a balance between goodness of fit and parsimony, with the two not always being positively related (Myung, Balasubramanian, & Pitt, 2000).

In light of this, this article highlights and discusses some of these considerations consisting of three parts: (a) Provide an overview of six statistical or psychometric issues and approaches to be aware of, investigate, and report when

CONTACT Thomas A. Schmitt tom@equastat.com Equastat, LLC, Technology Innovation Center, 10437 W Innovation Dr, Milwaukee, WI 53226.

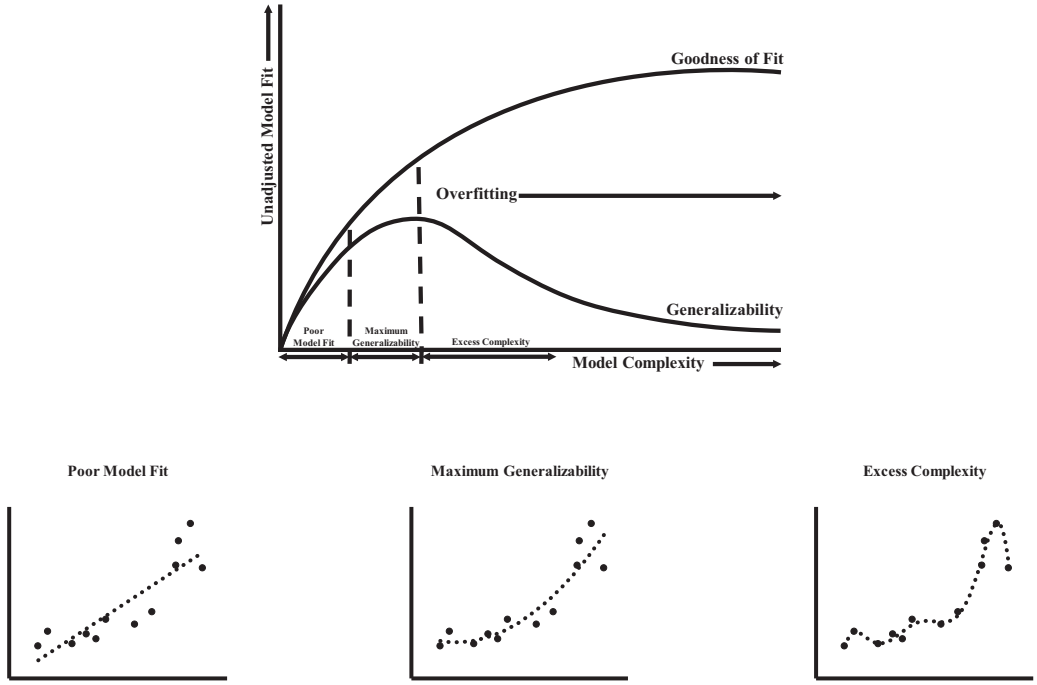

Figure 1. An illustration of the relationship between generalizability, goodness of fit, and overfitting as a function of model complexity (Myung & Pitt, 2002). There is a point of maximum generalizability or maximum predictive validity after which models can exhibit excessive complexity and a decrease in generalizability due to overfitting the data. This is illustrated in the three models below the plot.
Source: From *Stevens' Handbook of Experimental Psychology* (p. 449, Figure 11.4), by J. Wixted (Ed.), 2002, New York, NY: Wiley. Copyright 2002, with permission from Wiley.

engaging in factor structure validation, along with a flowchart for recommended procedures to understand latent factor structures (Figure 2); (b) by way of demonstrating these matters, we provide a summary of the updated Posttraumatic Stress Disorder Checklist (PCL–5[1]; Weathers, Litz, Herman, Huska, & Keane, 1993; Weathers et al., 2013) factor models (see Table 1) and a rationale for validation; and (c) we conduct a comprehensive statistical and psychometric validation of the PCL–5 factor structure to demonstrate all issues and work through the Figure 2 flowchart. We believe this will provide researchers and practitioners a very practical and actionable process for engaging in measure and factor structure validation, also providing a complete and through validation of the PCL–5.

It is important to note that our goal is not to provide an exhaustive review of factor model evaluation methods (see Schmitt, 2011, for a review), but to review and illustrate factor model selection issues and methods through a comprehensive validation for a widely researched and commonly used diagnostic instrument, the PCL–5. This means the methods discussed, recommended, and illustrated here using the PCL–5 are broadly applicable and can be applied to most any instrument development and validation process that involves psychometric factor dimension reduction and model selection.

Six statistical and psychometric issues and approaches

From a statistical and psychometric perspective, researchers should be cognizant of six issues and approaches: (a) item and variable skew, (b) model estimation, (c) factor analysis frameworks (i.e., implementing exploratory factor analysis [EFA], confirmatory factor analysis [CFA], or both; see Figure 2), (d) number of items per factor, (e) interfactor correlation magnitudes, and (f) factor model selection. Each of these should be evaluated and at least briefly discussed, with the appropriate statistics reported. Although all these areas are critical, perhaps most important is selecting the best factor structure. To help guide researchers, Figure 2 outlines a potential methodology to explore the factor structure under different modeling situations.

Item and variable skew

As with any statistical analysis, researchers should carefully examine the skewness and other descriptive statistics to determine the most statistically appropriate estimation method and modeling approach. Examining item skewness, in particular, is essential because extreme item distributions can affect the estimation accuracy (Flora & Curran, 2004), the item's intercept or threshold (depending on the estimation method used) estimates, the correlation between the variables, and, consequently, the dimensional structure. Specifically, skewed items with similar thresholds can generate spurious factors, called *difficulty factors* or *method factors* (Bernstein & Teng, 1989; Coenders, Satorra, & Saris, 1997; McDonald & Ahlawat, 1974).

As demonstrated by the few items per factor in the PCL–5, skewed items can cluster as difficulty factors (e.g., items associated with very rare psychological occurrences are likely to correlate highly due to their similar item distribution), rather than sharing a similar construct domain. This occurrence can also result in unstable and conflicting factor models, as the data skew (rather than the item content) and sample characteristics can drive a factor structure. Moreover, these skewed items can

[1]The PCL was also updated to the 20-item PCL–5 (Weathers et al., 2013) to match the 20 posttraumatic stress disorder (PTSD) symptoms of the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* ([*DSM–5*]; American Psychiatric Association, 2013; see Bovin et al., 2016).

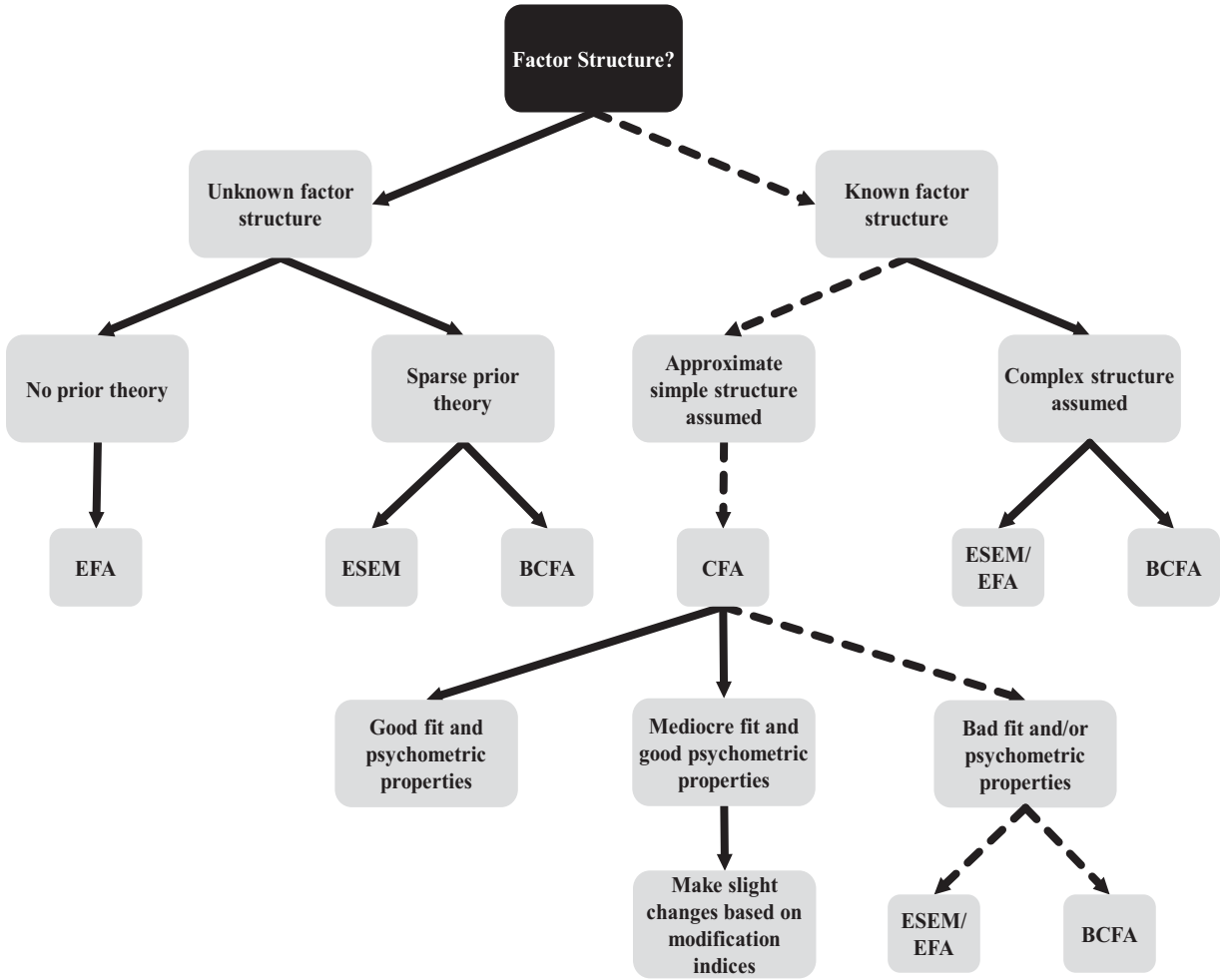

Figure 2. A flowchart of recommended procedures to understand latent factor structures. We encourage researchers to always conduct dimensionality analyses with their CFA, BCFA, EFA, or ESEM to ensure the correct number of factors are estimated and to understand any potential model misfit. This diagram is not all inclusive and researchers should consider other modeling issues as appropriate. The dashed lines represent the procedure followed in our analysis of the updated Posttraumatic Stress Disorder Checklist (PCL–5). *Note*. If factors are highly correlated, a large dominant eigenvalue is present, and theory or the analyses (i.e., CFA, BCFA, EFA, and ESEM) suggest a general dominant factor, then testing a bifactor model would be appropriate. CFA = confirmatory factor analysis; BCFA = Bayesian confirmatory factor analysis; EFA = exploratory factor analysis; ESEM = exploratory structural equation modeling.

negatively influence the model fit (McLeod, Swygert, & Thissen, 2001; Sawaki, Stricker, & Oranje, 2009), which might also lead to inconsistent results across samples. Considering the results from Table 1, it is concerning that only two authors reported or tested the distribution properties of the data.[2]

Model estimation

Depending on the degree of item skew and response option proliferation, it is critical that researchers employ appropriate model and estimation methods (Flora & Curran, 2004; Wirth & Edwards, 2007). Fortunately, more robust linear and nonlinear factor analysis estimation methods and frameworks exist that authors can employ to circumvent these data issues (e.g., weighted least squares mean- and variance-adjusted [WLSMV], robust maximum likelihood [MLR], and Bayesian). Nonetheless, researchers need to provide a valid justification for selecting the estimation method(s) that are based on previous statistical research, rather than user preference. It is also important that researchers use consistent estimation approaches across factor analysis methods (e.g., EFAs and CFAs) so as not to introduce estimation method variation into their study. For example, if authors use maximum likelihood (ML) to estimate their EFA model and WLSMV to estimate their CFA model (especially with the same sample), it makes it difficult to decipher whether the estimation method (e.g., ML vs. WLSMV) or modeling approach (i.e., EFA vs. CFA) created differences in model results.

As indicated in Table 1, most authors (66%) reported using either WLSMV or MLR. Although it is a positive finding that many authors used more robust methods, it is critical to keep in mind that WLSMV typically performs better than MLR (Li, 2016) with ordered categorical data. Other authors used ML estimation, which has been shown to perform poorly with skewed ordered categorical data, or simply did not report the estimation method. Even when using robust estimators (i.e., WLSMV or MLR) within factor analysis, it is still possible to have difficulty factors or spurious factors with highly skewed data or when items have large differences in thresholds (Yang & Xia, 2015). Thus, researchers should be exploring and considering the item distribution even when implementing

[2]The American Psychological Association provides reporting standards for quantitative research in psychology, which includes structural equation modeling standards in Table 7 (Appelbaum et al., 2018).

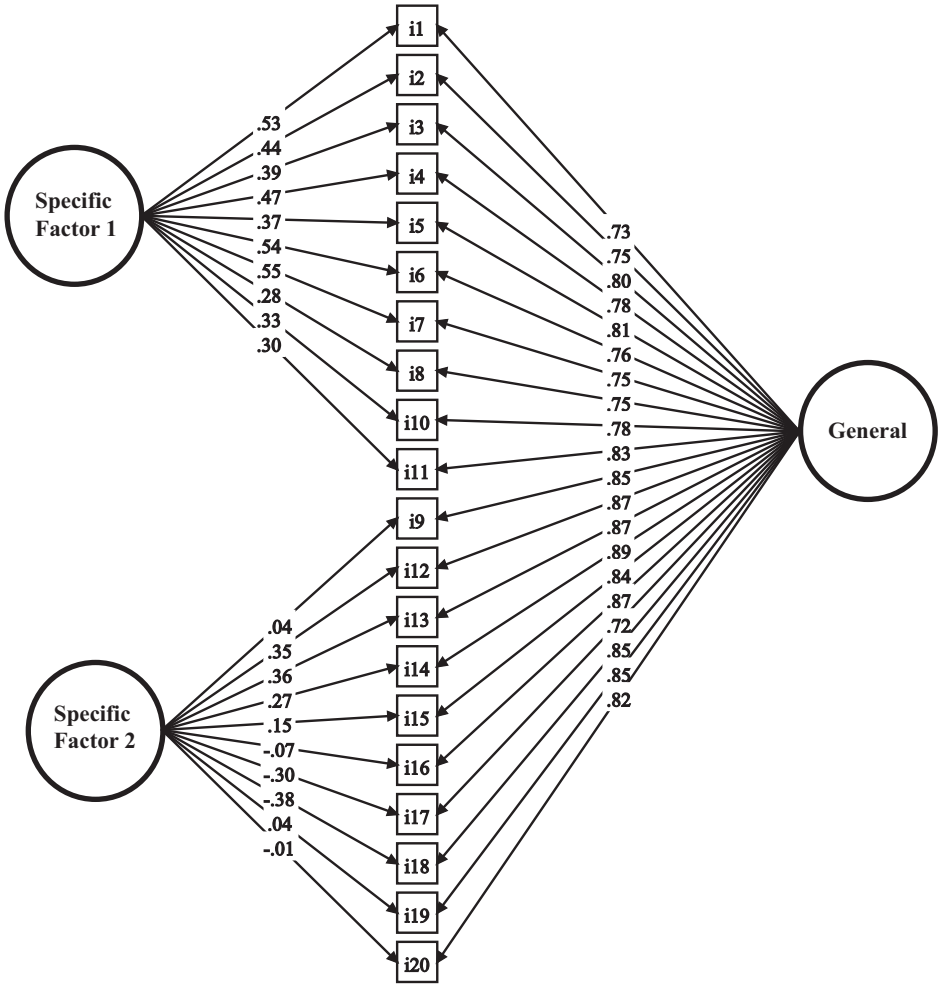

Figure 3. The factor model and standardized factor loadings for the two-factor bifactor model.

robust estimation methods. Given that a full literature review on factor analysis model estimation is outside the purview of this study (see Beauducel & Herzberg, 2006; Flora & Curran, 2004; Li, 2016; Muthén & Asparouhov, 2012), it is our hope that future researchers not only report and select the most appropriate estimation method, but also consider how the estimation method influences the parameter estimates.

Although some authors did use robust estimation methods (e.g., WLSMV), these methods still rely on using model modification indexes to free parameters in a stepwise fashion within CFA. Although freeing parameters in a stepwise fashion might seem logical, with each unconstrained parameter the model can stray further and further from the original theory with an overfit and incorrect model. Alternative methods do exist that might help to illuminate the latent factor structure and avoid these pitfalls, with Bayesian CFA (BCFA) being one such approach (Asparouhov, Muthén, & Morin, 2015; Muthén & Asparouhov, 2012). BCFA gives researchers the flexibility to identify potential cross-loadings, or residual covariances that might have otherwise been overlooked. If done properly (see Asparouhov et al., 2015), BCFA can result in more accurate factor correlations and helps to discover instances where the CFA model fails (e.g., relevant model modifications are discovered), while also avoiding model modifications from small and unimportant residual correlations.

Factor analysis frameworks

Another important consideration is whether to use an EFA, CFA, or a combination of both, including exploratory structural equation modeling (ESEM; Marsh, Liem, Martin, Morin, & Nagengast, 2011) or BCFA (Asparouhov et al., 2015; Muthén & Asparouhov, 2012). Unfortunately, some misconceptions exist around which, when, and how to use various factor analytic approaches. Researchers also typically hold the erroneous view that estimating an EFA and CFA model with the same data is inappropriate. Although it is often preferable to estimate the EFA and CFA model with different samples, especially for cross-validation, it is perfectly acceptable to fit different models to the same data in an effort to better understand the data generating process and factor structure. In fact, some sample exploration and confirmation should be encouraged under certain circumstances (e.g., poor CFA model fit), especially when assuming a priori theory and only using confirmatory models can led to statistically weak, complex, and overfit models (Browne, 2001; Gorsuch, 1997).

As opposed to blindly using CFA for rote factor structure confirmation and covert exploration (Asparouhov & Muthén, 2009; Browne, 2001; Gorsuch, 1983; MacCallum, Rosnowski, & Necowitz, 1992; Ropovik, 2015), it is arguably better to assess and reassess the data with an EFA to find where and why model

Table 1. Summary of factor analytics studies for the updated Posttraumatic Stress Disorder Checklist (PCL–5).

First author	Year	Sample size (*n*)	Sample demographic	Item/variable skew discussed	Model estimation method	Factor analysis method	Interfactor correlations reported	Factor selection method(s)	Tested/best PCL–5 model(s)
Armour	2016	412	U.S. fully trauma-exposed undergraduate students	No	WLSMV	CFA	Yes (.61–.89)	1, 2, 3, 4, 5, 9	2, 3, 4, 5, 6, **7**
Armour	2015	497 & 1,484	U.S. fully trauma-exposed undergraduate students and veterans	No	WLSMV	CFA	Yes (.59–.92)	1, 2, 3, 4, 5, 6, 9	2, 3, 4, 5, 6, **7**
Ashbaugh	2016	838 & 262	Canada fully trauma-exposed English/French speaking undergraduate students	No	MLE	CFA	Yes (.60–.95)	1, 2, 3, 5, 6, 7, 8	2, 5, **7**
Biehn	2013	266	U.S. fully trauma-exposed undergraduate psychology research subjects	No	WLSMV	CFA	Incomplete (.71–.91)	1, 3, 4, 5, 9	**2, 3**
Blevins	2015	278 & 558	U.S. fully trauma-exposed undergraduate students	No	MLR	CFA	No	1, 2, 3, 4, 5, 6, 7, 8, 9	2, **5**, 7
Bovin	2016	328 & 140	U.S. fully/partially trauma-exposed veterans	Yes	MLE	CFA	No	1, 3, 4, 5, 6, 7, 8, 9	1, 2, 6, **7, 8**, 9, 10, 11
Eddinger	2017	129 & 737	U.S. partially trauma-exposed veterans and undergraduate students	No	MLE	CFA	No	1, 3, 4, 5, 6, 8	2, 3, **4**, 9
Frewen	2015	557	U.S. PCL–5 PTSD diagnosed community sample	No	PAF/MLR	EFA/LPA	Yes[a] (–.31–.31)	10, see Frewen for LPA	EFA 6-factor and LPA 5-class best
Keane	2014	507	U.S. stress-exposed veterans	Yes (not reported)	MLR	CFA	Yes (.79–1.31)[b]	1, 2, 3, 4, 5, 7, 9	**2**
Krüger-Gottschalk	2017	352	German fully trauma-exposed community sample	No	MLR/WLSMV	CFA	Yes (.76–.93)	1, 2, 3, 4, 5, 6, 7, 8, 9	2, **3**, 4, 5, 6, 7
Liu	2014	1,196	Chinese fully trauma-exposed earthquake survivors	Yes (limited)	MLR	CFA	Yes (.60–.97)	1, 2, 3, 4, 5, 6, 7, 8, 9	2, 3, 4, 12, 13, **14**
McSweeney	2016	290: 113 + 177	U.S. partially trauma-exposed undergraduate students/ Amazon Mechanical Turk	No	Not reported	EFA	No	1, 3, 4, 5	5 models tested
Mordeno	2016	460	Filipino partially trauma-exposed hurricane survivors	Yes (limited)	MLR	CFA	Yes (.42–.85)	1, 2, 3, 4, 5, 6, 7, 9	5-factor best 2, 3, 4, 5, 6, **7**
Murphy	2017	364 or 481 (*n* unclear)	Malaysian partially trauma-exposed adolescent community sample	No	MLR/WLSMV	CFA	No	1, 3, 4, 5, 6, 7, 8	2, 3, 4, 5, **6, 7**, 16
Pietrzak	2015	1,484	U.S. partially trauma-exposed veterans	No	Not reported	CFA	No	Not reported	2, 5, 6, **7**
Shevlin	2017	195 & 239	U.K. fully trauma-exposed clinical samples	No	MLR/WLSMV	CFA	No	1, 3, 4, 5, 6, 7, 8	2, 3, 4, **5**, 6, **7**, 16
Tsai	2015	1,484	U.S. partially trauma-exposed veterans	No	MLR	CFA	No	1, 2, 3, 4, 5, 6, 7, 8, 9	2, 4, **15**
Wortmann	2016	912	U.S. treatment-seeking veterans and military personnel	No	Not reported	CFA	No	1, 2, 3, 4, 5, 7, 8, 9	2, 4, 5, 6, **7**

Note. Factor selection method: 1 = χ^2; 2 = $\Delta\chi^2$; 3 = comparative fit index; 4 = Tucker–Lewis Index; 5 = root mean square error of approximation (RMSEA); 6 = confidence interval for RMSEA; 7 = standardized root mean square residual or weighted root mean square residual; 8 = Akaike's information criterion; 9 = Bayesian information criterion; 10 = eigenvalue > 1 criteria; 11 = parallel analysis; 12 = misspecification examination. Tested/best PCL–5 model(s): 1 = posttraumatic stress disorder (PTSD) one-factor; 2 = *DSM–5* four-factor; 3 = *DSM–5* dysphoria four-factor; 4 = *DSM–5* dysphoric arousal five-factor; 5 = anhedonia six-factor; 6 = externalizing six-factor; 7 = hybrid seven-factor; 8 = anhedonia seven-factor; 9 = *DSM–4* three-factor; 10 = *DSM–4* Dysphoria four-factor; 11 = *DSM–4* dysphoria arousal four-factor; 12 = *DSM–5* revised five-factor; 13 = *DSM–5* dysphoria revised five-factor; 14 = *DSM–5* dysphoric arousal revised six-factor; 15 = *DSM–5* new model six-factor; 16 = alternative dysphoria sic-factor. The best model is shown in bold. WLSMV = weighted least squares mean- and variance-adjusted; CFA = confirmatory factor analysis; MLE = maximum likelihood estimation; MLR = robust maximum likelihood; PAF = principal axis factoring; EFA = exploratory factor analysis; LPA = latent profile analysis.

[a]Used LPA to estimate factor correlations.

[b]Correlations above one can indicate model misspecification.

misspecification is occurring. This is especially true when CFA model respecification is unsupported by theory or when poorly fitting CFA models result in a large number of modification indexes (e.g., Taylor & Pastor, 2007). Even with previously strong theory and having a good fitting CFA model, it is perfectly reasonable (and perhaps very wise) to either precede or follow up a poor fitting CFA model with an EFA to better understand the factor structure. Furthermore, it can be fruitful to conduct a CFA on the same data used with an EFA. This is especially true when the researcher seeks to demonstrate the model's fit under a more restrictive model, allow for the comparability of CFA model fit with previous and future research, and establish the degree to which the results (e.g., interfactor correlation) differ based on model specification (see Sass & Schmitt, 2010).

Rather than following a scientifically limiting approach to model building, researchers should concentrate on estimating the correct or best model within a confirmatory framework, exploratory framework, or both, which is conceptually supported by the integration of CFA and EFA in the SEM framework of ESEM (Asparouhov & Muthén, 2009). This is especially important, because exploring complex structures with a CFA can lead to rather arbitrary modifications, unrealistic factor loadings, and elevated interfactor correlations (Marsh et al.,

2009; Schmitt & Sass, 2011). Within ESEM, researchers can incorporate an EFA and CFA simultaneously (e.g., two factors might be represented as an EFA and one as a CFA) to better represent their data. As Conway and Huffcutt (2003) found, researchers make better decisions when EFA plays a more consequential role in their research to develop and test theoretical models, refine previously developed instruments, and test the factor structure's generalizability (Yang & Xia, 2015). In summary, EFA can be used to explore poorly fitting CFA models, test factor structures that lack strong theory or hypotheses, and confirm theorized factor structures when the CFA independent cluster assumption is unrealistic (Gorsuch, 2003; Schmitt, 2011).

In reviewing Table 1, most PCL–5 authors used a CFA, which seems appropriate at first glance given that the PCL has been extensively researched and the PCL–5 is theoretically grounded in the *DSM–5*. With that said, many authors used CFA in a more exploratory fashion by just simply adding factors to improve global fit or relying on modification indexes, which is arguably inappropriate (e.g., Asparouhov & Muthén, 2009; Byrne, 2012; Gerbing & Hamilton, 1996; Gorsuch, 1983; Hayduk, 2014b; MacCallum et al., 1992; Marsh et al., 2011; Mulaik, 1972). In cases like this, using EFA and CFA together is more rigorous, because EFA, along with dimensionality analyses (e.g., parallel analysis and eigenvalues), encourages a more holistic and flexible statistical approach to item and factor evaluation and facilitates a better understanding of the optimal factor structure.

Another factor analytic approach that lends itself well to the PCL–5 and many other measures is the *bifactor* model. Bifactor models empower researchers to provide evidence of a single general factor, as well the multidimensionality through parcels (or doublets) of items that are not necessarily strong factors, but simply groupings of items that represent similar content domains (Reise, Moore, & Haviland, 2010). The bifactor model also lets the general factor and subfactors (i.e., specific factors) to "compete" against one another in explaining item response variability. Thus, a bifactor model can demonstrate evidence of unidimensionality, where there is a unidimensional general factor and weak to nonexistent subscale factors. Moreover, bifactor models can also inform the researcher whether these subscale factors are (a) really just item parcels, (b) a multidimensional model with strong subscale factors and a weak general factor, or (c) a balance of equally strong general factor and subscale factors. It is worth noting that bifactor models only have first-order factors, which is distinct from second-order factor models (Canivez, 2016; Reise, 2012).

In general, the challenge of having a single general factor with evidence of other interpretable factors and improved fit alleviates researchers from having to postulate and fit numerous multidimensional factor models (Reise et al., 2010). Further, it overcomes the ambiguity of the general practice of total scores, and not taking into account subscales, for clinical diagnosis when additional item groupings have been found. Bifactor models can also be useful for modeling method effects, such as those that result from negatively and positively worded items.

It is important to emphasize that another strength of the bifactor model is in testing for multidimensionality and examining the bifactor indexes (see Reise, Bonifay, & Haviland, 2013b, and Rodriguez, Reise, & Haviland, 2016, for a review of bifactor indexes). The bifactor indexes are discussed briefly later, but Canivez (2016) and Rodriguez et al. (2016) provided a recent tutorial of this model for interested readers. Using the bifactor model indexes, researchers can answer five basic questions:

1. Do the data represent a unidimensional or multidimensional factor structure?
2. Is it appropriate to use the overall or total score rather than the individual or subscale?
3. Does the total score represent a single reliable latent construct and are the subscales still be reliable after adjusting for the general factor?
4. Are subscale scores truly independent of the general factor?
5. How well do a set of items depict the latent construct and can these items be used to specify latent constructs for SEM?

Items per factor

As Reise, Waller, and Comrey (2000) noted when revising scales, researchers should carefully consider the number of factors that exist in the data, the correlation between these factors (discussed later), and the number of items per factor, with the understanding that it is preferable to increase reliability by having more items (i.e., overinclusive) than less (i.e., underinclusive). Unfortunately, item sparseness is a major problem on many of the theorized PCL–5 factor structures (i.e., many factors have only two items). The existence of too few items per factor becomes not only an issue of model identification and replication (Bollen, 1989; Little, Lindenberger, & Nesselroade, 1999; Velicer & Fava, 1998), but also a matter of construct underrepresentation (Kaplan & Saccuzzo, 2008).

Although it is rarely appropriate to have two-item factors (Raubenheimer, 2004), it is acceptable to have small item groupings known as facet scales (i.e., homogeneous item clusters or item parcels) that represent different aspects of a construct (Reise et al., 2000). Facets have high item correlations, and if combined, have better reliability and distributional properties than single items. As discussed later, this would also justify the large interfactor correlations in the PCL–5. Therefore, based on the opinion of psychometric and statistical experts, PCL–5 researchers should avoid two-item factors and instead conceptualize their utility as facets and, as discussed, consider fitting alternative models, such as bifactor models (Reise et al., 2010).

Interfactor correlation magnitudes

Farrell (2010) indicated that when high interfactor correlations exist (e.g., $|r| > .75$), concerns arise related to discriminant validity (i.e., the degree to which measures of different constructs or concepts are different). Not only is discriminant validity an issue from a psychometric perspective, but trepidations related to multicollinearity within future regression-based models, including SEM and path analysis, will also surface. Large interfactor correlations can also result from overly restrictive CFA models that mask the true factor structure (Marsh et al., 2009; Schmitt & Sass, 2011). Unfortunately, these interfactor correlations are rarely provided or discussed within the PCL–5 validation research (see Table 1). Our results, and other PCL–5 studies (Armour, Műllerová, & Elhai, 2016;

Armour et al., 2015; Frewen, Brown, Steuwe, & Lanius, 2015; Keane et al., 2014; Liu et al., 2014), indicated the interfactor correlations for many of the PCL–5 factors are large, suggesting poor discriminant validity.

Associated with dimensionality (i.e., the number of factors extracted) and discriminant validity, it is worth addressing a scenario where researchers might argue for additional factors beyond what the data recommend and with rather high interfactor correlations. As Preacher et al. (2013) discussed, if overfactoring produces a more stable factor model and better predicts the desired outcome, then these facets could be worth treating as factors (still assuming the number of items on that factor was acceptable). As an example, researchers could conceivably argue that a five-factor PCL–5 model that better predicts (e.g., has an $R^2 = 0.65$) a desired outcome (e.g., suicide tendencies) is preferable to a one-factor model with poorer model predictions (e.g., $R^2 = 0.45$). With that said, there has been no research that provides evidence that these more complex models predict significantly better with the PCL–5.

Factor selection

One of the most important and difficult challenges in factor analysis is determining the "correct" number of factors, which is strongly connected to whether a CFA or EFA model is fit. It is not surprising that choosing the right solution is a difficult undertaking. Especially when theory is weak, the measure has moderate to weak validity and reliability, item distributions are not ideal (i.e., not normally distributed or insufficient variability), and subjects are not randomly selected or representative (Cudeck & Henly, 1991). Unfortunately, the published PCL–5 studies have not used additional approaches (e.g., parallel analysis) to determine the number of factors, as most (see Table 1) relied solely on global fit evaluation with the χ^2 test, approximate fit indexes (AFIs), or the likelihood ratio test ($\Delta\chi^2$) to compare nested models.

Weaknesses (e.g., sensitivity to large sample sizes, model complexity, nonnormal data, and model misspecification) associated with the χ^2 and $\Delta\chi^2$ have been well documented (for a review, see Herzog, Boomsma, & Reinecke, 2007; Hoyle, 1995; Hu & Bentler, 1998; Sass, Schmitt, & Marsh, 2014). These weaknesses result in the tendency of the χ^2 and $\Delta\chi^2$ to reject correctly specified models under many of the conditions associated with the PCL–5. More specifically, this means the χ^2 statistic has a tendency for factor overextraction (i.e., the number of extracted factors is larger than the true number of factors; see Asparouhov & Muthén, 2009; Hayashi, Bentler, & Yuan, 2007), which then results in the $\Delta\chi^2$ statistic no longer following a χ^2 distribution and again producing biased inferences (Geweke & Singleton, 1980).

Due to these limitations, many researchers often rely heavily on the AFIs, such as the Tucker–Lewis Index (TLI), comparative fit index (CFI), standardized root mean square residual (SRMSR) or weighted root mean square residual, and root mean square error of approximation (RMSEA). Despite the development of AFIs to circumvent problems associated with the χ^2 statistic, model fit can still be affected by model misspecifications (recall, many of the AFIs are a function of the χ^2) and other model characteristics (Saris, Satorra, & van der Veld, 2009). Thus, the χ^2 and AFIs should be used judiciously and in conjunction with other methods of factor extraction. For example, authors might evaluate the parallel analysis, eigenvalues, χ^2 test, and AFIs to provide an initial assessment of the number of factors, and then use the $\Delta\chi^2$ test and ΔAFIs to determine whether a model with fewer or more factors is preferable. This approach might provide some indication of improvement in a more complex model over a more parsimonious model.

Despite the numerous alternatives to determining the number of factors, there exists no single best statistical criterion for doing so (Gorsuch, 2003), which means researchers must examine the results using different methods and decide the number of factors in an integrated fashion (Hayashi et al., 2007). Researchers must also provide strong evidence and justification that are supported by data and theory when finalizing the optimal number of factors.

Although given less attention here, many PCL–5 researchers relied on Akaike's information criterion (AIC) and Bayesian information criterion (BIC), along with the change in these statistics, to determine the number of factors. Although these methods can be regarded as single sample estimates of an expected cross-validation criterion and certainly have a place in statistics (see Preacher & Merkle, 2012; Preacher et al., 2013), their weaknesses are also well documented. Mulaik (2009) criticized the AIC and BIC due to their poor performance, sample size dependency, and the frequency with which they are misunderstood or misapplied. Katz (1981), Shibata (1976), and Preacher and Merkle (2012) further indicated that the AIC and BIC tend to overestimate the number of parameters needed and favor model complexity as sample size increases (Bozdogan, 2000; McDonald & Marsh, 1990; Mulaik, 2001). Further, AIC and BIC are only indirect estimates of generalizability and make more assumptions about the underlying model than direct cross-validation methods (see Hastie et al., 2009; James et al., 2013). The one major advantage that the AIC and BIC methods have over other model fit statistics is their ability to compare nonnested models (i.e., allows for a comparison of models with a different number of variables); however, given that the PCL–5 models are always nested, this should not be the rationale for their use.

Fortunately, methods exist that when used in combination can increase model accuracy, statistical rigor, and practicality, and can arrive at a reasonable number of factors that are psychometrically sound (see Schmitt, 2011). Considering this, we provide a potential roadmap (see Figure 2) for researchers to follow when evaluating the factor structure. In the context of Figure 2, researchers need to know when and how to do the following:

1. Conduct a parallel analysis and consider the eigenvalue magnitudes to provide an initial assessment of the preliminary number of factors.
2. Implement an EFA, ESEM, BCFA, or bifactor model to explore the factor structure (i.e., evaluate the interfactor correlations, along with the primary and secondary factor loading magnitudes) and evaluate the appropriateness of the latent constructs.
3. Conduct a CFA to determine the model quality under more restrictive conditions and determine whether the parameter estimates change significantly.
4. Evaluate the EFA and CFA models with the χ^2 test, AFIs, and $\Delta\chi^2$ test.

5. Examine model misspecification with the Saris–Satorra–van der Veld method to complement χ^2 and AFI statistics. This last point can be especially important because as Hayduk (2014a, 2014b) lucidly summarized, when the χ^2 depicts covariance ill fit, potentially serious model misspecifications should be thoroughly investigated.

It is worth noting that other factor extraction methods exist (see Peres-Neto, Jackson, & Somers, 2005; Revelle, n.d.; Schmitt, 2011), but they are not all accurate, statistically rigorous, or practical.[3] Regardless of the methods used, considerable attention must be paid to the number of factors estimated and the approach taken to arrive at the factor structure. Unfortunately, nearly all the PCL–5 studies to date rely almost exclusively on model selection based on the theorized number of factors and the χ^2 test, which appears to have led to a proliferation of complex factor models (see Table 1).

Summary of the PCL–5 factor models and rationale for validation

The updated PCL–5 (Weathers et al., 1993; Weathers et al., 2013) is an example of a measure with a factor structure that has been extensively investigated and hypothesized to exhibit excessively complex factor structures due to overfitting the data. The PCL–5 was selected as a demonstration measure for several additional substantive and psychometric or statistical reasons. First, the PCL–5 is widely used in many settings to make "life-changing" assessments, so it is a relatively high-stakes assessment (see Gray-Little & Kaplan, 1998; Padilla & Borsato, 2008). Second, although complex factor structures of the PCL–5 have been proposed, many of these factor structures are difficult to defend statistically and psychometrically. Nevertheless, determining the optimal theoretically and statistically derived factor structure based on the underlying symptom dimensions of PTSD has remained elusive. This discussion began with the *DSM–IV–TR* (American Psychiatric Association, 2000) and has persisted with the *DSM–5* (e.g., Armour et al., 2016; Blevins, Weathers, Davis, Witte, & Domino, 2015; Bovin et al., 2016; Konecky, Meyer, Kimbrel, & Morissette, 2016; Liu et al., 2014; Tsai et al., 2015; Wortmann et al., 2016). Despite this ongoing examination of PCL–5 factor structures, these models have not been evaluated from a more comprehensive and psychometric and statistical perspective. Last, and perhaps most important, in the context of personality assessment those diagnosed with PTSD have been shown to have a high rate of "character" pathology (e.g., Southwick, Yehuda, & Giller, 1993) and PTSD has strong connections with different types of personality disorders (e.g., Davidson & Foa, 1991; King, North, Surís, & Smith, 2016).

Consequently, the PCL–5 research has introduced at least 15 different one-, three-, four-, five-, six-, and seven-factor models, ranging from the one-factor PTSD model to the seven-factor hybrid model (see Table 1; Armour et al., 2016; Bovin et al., 2016; Young, 2016). The models most often chosen as "best" by researchers tend to be the more complex models, which generally have encompassed the six-factor anhedonia, six-factor externalizing behaviors, and seven-factor hybrid (i.e., a hybrid of the externalizing behaviors and anhedonia models). These more complex models have generally been selected as best based largely on theory driven by global model fit criteria, often ignoring other psychometric and statistical methods. Although theoretical justification is extremely important when deciding on the final factor structures, overreliance on global fit indexes can hinder scientifically reproducible results (Baker, 2016; Open Science Collaboration, 2015; Ropovik, 2015; Wasserstein & Lazar, 2016) for determining the best factor structure around practical and clinical efficacy.

In the context of the previous paragraph, Table 1 provides a summary of PCL–5 CFA and EFA studies to date and the statistical methods used. Several patterns are worth highlighting. First, when determining the optimal number of factors, researchers adhered mostly to CFA models and model fit statistics, ignoring the eigenvalues (i.e., the eigenvalues were not provided nor discussed) and related methods (e.g., parallel analyses) to aide in determining the number of factors. Second, most authors argued for more complex models without considering if these item clusters truly are factors rather than simply facets (see Reise et al., 2000). This is problematic because the artifact of these complex models can be statistically dubious factors with a small number of items per factor (i.e., often only two) that are not psychometrically sound latent constructs. This can lead to excessive complexity and falling into the overfit trap, as discussed and illustrated in Figure 1. Finally, the interfactor correlations were rarely provided; thus, there was no evidence of discriminant validity and whether the factors should or even could be combined. For those studies that did provide the interfactor correlations, these correlations were often so large it would be difficult to argue different factors were being measured.

Statistical and psychometric validation of the PCL–5 factor structure

Participants

This study uses responses from U.S. Air Force MQ-1 Predator and MQ-9 Reaper remotely piloted aircraft (RPA; i.e., drones) pilots and sensor operators, who have high levels of direct exposure to combat-related trauma. Specifically, RPA pilots and sensor operators actively track, target, and destroy enemy combatants and assets; offer protection to civilian and military personnel; inspect and survey battle damages after they have performed weapon strikes (e.g., Hellfire missile strikes); and obtain information to increase situational awareness of the battlefield. Resulting from their service, RPA pilots and sensor operators can suffer PTSD or severe trauma due to witnessing death and destruction from the weapon strikes they perform, as well real-time observation of the battlefield (Chappelle, Goodman, Reardon, & Thompson, 2014). The PCL–5 is used to help psychologists screen for and better understand the symptomology of trauma within this population and to track their level of PTSD symptoms over time.

[3] The eigenvalue-greater-than-1 rule or Kaiser criterion (K1) somehow still manages to persist in practice, but it should not be used as it has been shown to be inaccurate. Van der Eijk and Rose (2015) provided a nice open access review of the problems with the K1 criterion.

The purpose and methodology of the study were reviewed and granted exemption by the U.S. Air Force Research Laboratory Institutional Review Board and were considered to be minimal risk. Before participants began the electronic survey, they were asked if they understood the nature, purpose, and instructions of the survey and then were asked to voluntarily consent to participate. Those who endorsed "yes" proceeded to take the survey, whereas those who endorsed "no" were not given the survey and instead were redirected to another Web page that instructed them on how to contact the independent study researchers for additional information. Seven individuals declined participation after reading the informed consent section.

Sample 1

Data were collected from 1,403 RPA military pilots between January and April 2015. Most of the pilots were male (87.5% male, 12.4% female, 0.6% missing) and were either married (64.8%, $n = 909$) or single, never married (22.0%, $n = 309$). However, many were single due to divorce (7.1%, $n = 99$); unmarried, but in a significant or partner relationship (6.0%, $n = 84$); or did not answer this item (missing; 0.1%, $n = 2$). The age breakdowns were as follows: 18 to 25, 17.4% ($n = 246$), 26 to 30, 27.2% ($n = 381$), 31 to 35, 25.9% ($n = 363$), 36 to 40, 15.3% ($n = 215$), and older than 41, 13.6% ($n = 191$). As for military experience, most were currently classified as active duty (66.4%, $n = 931$) or in the Guard (27.4%, $n = 384$), with a significantly smaller number classified as Reserve (5.6%, $n = 78$) or civilian, contractor, or missing (0.7%, $n = 10$).

Sample 2

Whereas Sample 1 included all RPA pilots regardless of their level of trauma, Sample 2 was selected to determine if the results differed when using a more traumatized sample. Therefore, a subsample of participants from Sample 1 were selected who were directly tasked with either decision making or analyzing real-time audio or video feed from the RPA strikes. To be included in Sample 2 ($n = 715$), participants were required in the past 12 months (at the time of data collection) to have one or more severe events on one of the following questions:

1. In total, how many separate events involving U.S. or Allied Forces being physically injured or killed by enemy forces have you virtually observed?
2. In total, how many separate events involving U.S. or Allied Forces being injured or killed by friendly forces have you virtually observed?
3. In total, how many separate events involving noncombatant bystanders being injured or killed as a result of enemy forces operations have you virtually observed?

It is critical to indicate that not all results were provided for Sample 2, as these results were mostly identical to Sample 1. Omitted results from Sample 2 can be obtained from the corresponding author.

Measures

The PCL–5 is a 20-item screening instrument based on *DSM–5* symptom Criteria B (re-experiencing), C (avoidance), D (negative cognitions/mood), and E (arousal) clusters for PTSD (American Psychiatric Association, 2013; Weathers et al., 2013). Participants report the severity of symptoms over the past month they are currently experiencing in relation to a trauma-related event or exposure. Respondents rate each item on a scale from 0 (*not at all*) to 4 (*extremely*). A total symptom severity score ranges from 0 to 80 and can be obtained by summing the scores from each of the 20 items. Although the PCL–5 is a relatively new measure, Hoge, Riviere, Wilk, Herrell, and Weathers (2014) found it performed equivalently to the PTSD Checklist-Specific in a study on U.S. soldiers. However, as indicated earlier, a statistically rigorous factor structure is currently in question, and, therefore, clear operational definitions have yet to be created.

Procedures

Participation to complete the PCL–5 was encouraged by U.S. Air Force MQ-1 Predator and MQ-9 Reaper leadership (wing, group, squadron commanders) via a mass e-mail invitation to approximately 2,500 U.S. Air Force RPA pilots and sensor operators (from 23 separate squadrons within the continental United States) with government e-mail. These operators engage in around-the-clock missions that involve surveillance of various battlefields throughout the globe, as well as tracking and eliminating enemy combatants via weapon strikes. As a result, this unique group of military personnel has high levels of exposure to battlefield trauma (Chappelle et al., 2014). The e-mail explained the study purpose and confidentiality safeguards to maximize participation and self-disclosure. Interested participants were directed to a secure Web site to complete a consent form, demographics questionnaire, the PCL–5, and other mental health screening instruments. On average, it took participants 25 to 30 min to complete. Participants were also instructed on local resources and points of contact for obtaining mental health care at their discretion.

Model estimation

All primary EFAs and CFAs were conducted within M*plus* 8 (Muthén & Muthén, 1998–2017) using a WLSMV estimator and a polychoric correlation matrix designed for ordered categorical data. WLSMV has been shown to perform equally well or better than other estimation methods with ordered categorical and skewed data (Flora & Curran, 2004; Liang & Yang, 2014). For our Sample 1 data, the items were all skewed and contained skew statistics that ranged from 1.41 (Item 20) to 4.45 (Item 16).

The latent factor variances were fixed at one to identify the model and set the metric for the CFA, BCFA, and bifactor models, whereas the mean and variance were fixed at 0 and 1, respectively, to identify the model for EFA models (i.e., typical EFA practice). Whereas the CFA factors were established by determining what items load on each factor using previous PCL–5 theory (see references earlier), EFA ascertained the factor structure by evaluating the eigenvalue magnitudes, parallel analysis, and model fit statistics. To obtain an approximate simple structure with EFA, an oblique Geomin rotation was used because the PCL–5 is a well-developed measure that should have fewer and smaller cross-loadings and produce a cleaner

Table 2. Model fit statistics for each of the previously theorized models using Sample 1.

Model	WLSMV model fit								MLR	
	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	TLI	RMSEA	RMSEA 90% CI	AIC	BIC
CFA and EFA one-factor	2167.27	170	239.34	2	0.951	0.945	0.092	[.088, .095]	46827	47142
CFA one-factor (modified model)	1843.39	168	CM	CM	0.959	0.953	0.084	[.081, .088]	46163	46488
CFA two-factor	1310.82	169	CP	CP	0.972	0.968	0.069	[.066, .073]	45228	45548
CFA four-factor *DSM–5*	1303.33	106	322.08	4	0.972	0.968	0.070	[.067, .074]	45073	45419
CFA four-factor dysphoria	1396.80	164	259.92	4	0.970	0.965	0.073	[.070, .077]	45137	45484
CFA five-factor dysphoria arousal	1194.32	160	426.28	8	0.975	0.970	0.068	[.064, .072]	44848	45215
CFA six-factor externalizing behaviors	1157.08	155	536.22	13	0.975	0.970	0.068	[.064, .072]	44717	45111
CFA six-factor anhedonia	684.08	155	562.85	13	0.987	0.984	0.049	[.046, .053]	44052	44446
CFA seven-factor hybrid	633.11	149	676.59	19	0.988	0.985	0.048	[.044, .052]	43914	44339

Note. All $\Delta\chi^2$ were statistically significant at the .001 level. WLSMV = weighted least squares mean- and variance-adjusted; MLR = robust maximum likelihood; CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root mean square error of approximation; CI = confidence interval; AIC = Akaike's information criterion; BIC = Bayesian information criterion; CFA = confirmatory factor analysis; EFA = exploratory factor analysis; CM = comparison model, which correlated two residual covariances (Item 7 with Item 6 and Item 17 with Item 18) and is the model that all other models were compared based on the $\Delta\chi^2$; CP = convergence problems. Although the results are not provided here, all models also differ significantly from the CFA two-factor model. When comparing the CFA one-factor (modified model) to the CFA two-factor model, this comparison had CP and, therefore, these results could not be obtained. Recall, WLSMV does not provide the AIC or BIC; thus, we reported these results using MLR estimation.

factor structure similar to CFA (see Sass & Schmitt, 2010; Schmitt & Sass, 2011).

Missing data

Although Little's missing completely at random test indicated the data were not missing completely at random with Sample 1, $\chi^2(520) = 769.052$, $p < .001$, the percentage of missing data was less than 1% (0.37%) and, therefore, treated using the default missing procedure in M*plus* (see Asparouhov & Muthén, 2010).

Model fit

Model fit was evaluated using the robust χ^2, CFI, TLI, and RMSEA. The χ^2 statistic is known to produce statistically significant values for good fitting models when the factor structure is complex and the sample size is large; thus, approximate model fit statistics (i.e., CFI, TLI, and RMSEA) were also evaluated. According to Hu and Bentler (1999), CFI and TLI statistics greater than 0.90 are deemed adequate and values greater than 0.95 are good. RMSEA values less than 0.10 and 0.06 are considered mediocre and good, respectively. More information about these statistics can be found in Hu and Bentler (1999). When comparing models, the DIFFTEST procedure was used to compute the $\Delta\chi^2$ and the ΔAIC and ΔBIC were considered. Recall that it is not statistically appropriate to compare the ΔCFI, ΔTLI, and ΔRMSEA with WLSMV estimation.

Techniques to better understand the latent factor structure

Confirmatory factor analyses

To replicate previous research (see Table 1), a series of WLSMV CFA models were first conducted. Although not commonly provided with previous PCL–5 psychometric research, a one- and two-factor WLSMV CFA model were also estimated using Sample 1 for comparability purposes. The one-factor WLSMV CFA produced a fairly good fitting model (see Table 2) and offered evidence of being a viable model (see Table 3). The two-factor WLSMV CFA model[4] (see Table 3) also fit the data well, but only slightly better than the one-factor WLSMV CFA model (see Table 2). Unfortunately, the interfactor correlation was extremely high (see Table 3), indicating low discriminant validity and high multicollinearity.

The model fit statistics for the six previously proposed PCL–5 factor structures (see Table 1) were provided in Table 2 with Sample 1 data. As expected, the inclusion of additional factors produced better fitting models based on the $\Delta\chi^2$ and AFI; however, one could argue that these models were overfitting the data (also evident based on the later dimensionality results) and do not fit the data significantly better from a practical standpoint. For example, although the ΔAFI statistics with WLSMV need to be interpreted with caution (Sass et al., 2014), these more complex models do not fit data much better than the one- or two-factor model based on the ΔAFI. Moreover, the interactor correlations were very high (nearly always above .80) and do not provide much evidence of discriminant validity.[5] From a modeling perspective, these factor structures should not be avoided in other statistical models (e.g., SEM, path analysis, multiple regression) due to significant multicollinearity concerns.

An alternative approach used by many PCL–5 researchers is to determine whether the "best model" fits significantly better than alternative models based on the $\Delta\chi^2$, ΔAIC, or ΔBIC. The argument for the $\Delta\chi^2$ is that if one model fits significantly better (based on some predetermined α level) than another model, this model is preferable. As indicated earlier, this approach has been extremely controversial among statisticians and has been shown not to always provide the best model, especially with large samples and complex models. Also used, and controversial among statisticians, are the ΔAIC and ΔBIC to select the

[4] Consistent with the procedures outlined in Figure 2, WLSMV EFA was conducted later because of less than ideal model fit and psychometric properties for WLSMV CFA. Because the EFA suggested either a one- or two-factor model, more parsimonious WLSMV CFA models were also estimated here for comparability purposes.

[5] All interfactor correlations corresponding to the models in Table 2 are available from the corresponding author.

Table 3. Weighted least squares mean- and variance-adjusted EFA and standardized CFA factor loadings for the one- and two-factor models using Sample 1 (n = 1,403) and Sample 2 (n = 715).

Item	One-factor (n = 1,403)	One-factor (n = 715)	Two-factor EFA (n = 1,403)		Two-factor EFA (n = 715)		Two-factor CFA (n = 1,403)		Two-factor CFA (n = 715)	
1	**.76**	**.86**	**.92**	−.03	**1.01**	−.14	**.89**		**.87**	
2	**.72**	**.85**	**.81**	.08	**.88**	−.01	**.87**		**.87**	
3	**.74**	**.88**	**.75**	.17	**.79**	.14	**.90**		**.90**	
4	**.81**	**.89**	**.83**	.10	**.75**	.18	**.91**		**.91**	
5	**.79**	**.87**	**.69**	.23	**.69**	.24	**.90**		**.89**	
6	**.80**	**.90**	**.94**	−.01	**.93**	.00	**.92**		**.91**	
7	**.76**	**.90**	**.95**	−.03	**.93**	.00	**.91**		**.91**	
8	**.64**	**.77**	**.56**	.28	**.54**	.29	**.81**		**.80**	
9	**.69**	**.81**	.38	**.50**	.38	**.49**		**.85**		**.85**
10	**.68**	**.82**	**.65**	.24	**.57**	.30	**.86**		**.84**	
11	**.77**	**.86**	**.62**	.31	**.58**	.34	**.90**		**.88**	
12	**.74**	**.87**	−.08	**.99**	.00	**.91**		**.91**		**.90**
13	**.76**	**.88**	−.11	**1.02**	−.06	**.97**		**.92**		**.90**
14	**.79**	**.90**	.00	**.93**	−.01	**.94**		**.92**		**.92**
15	**.72**	**.84**	.03	**.83**	.06	**.82**		**.85**		**.86**
16	**.64**	**.85**	.27	**.62**	.26	**.66**		**.87**		**.89**
17	**.52**	**.66**	.16	**.57**	.12	**.59**		**.71**		**.68**
18	**.64**	**.79**	.24	**.61**	.24	**.61**		**.83**		**.82**
19	**.70**	**.81**	.03	**.84**	.11	**.75**		**.85**		**.84**
20	**.65**	**.79**	.08	**.75**	.25	**.60**		**.82**		**.83**
r	N/A	N/A	.79		.77		.86		.86	

Note. EFA = exploratory factor analysis; CFA = confirmatory factor analysis. Factor loadings greater than .40 are shown in bold, with r representing the correlation between factors.

correct model, with change values typically greater than .10 being considered practically significant.

Not surprisingly (see Table 2), these more complex models consistently fit the data better based on the $\Delta\chi^2$, ΔBIC, and ΔAIC. Despite the $\Delta\chi^2$ always being statistically significant, this is arguably not enough evidence (especially due to the limitations associated with the $\Delta\chi^2$) to support these more complex models. The same argument could be made for the ΔBIC and ΔAIC. As emphasized later, it is unconvincing to claim these more complex models are psychometrically (especially when considering the few items per factor and the lack of discriminant validity) or statistically appropriate.

Exploratory factor analysis

Based on the WLSMV CFA results, we contend that no model is the definitive winner and the best model supported by the data from a statistically and psychometric perspective is inconclusive. To better understand the PCL–5 factor structure, we proceeded systematically through Figure 2, which is represented by the dashed line. The 20-item WLSMV EFA dimensionality results using Sample 1 revealed evidence of a single strong factor and a second weak factor based on the eigenvalues (first three eigenvalues were 14.10, 1.20, and 0.87) and a good fitting one-factor, $\chi^2(170) = 2167.27$, $p < .001$, CFI = 0.95, TLI = 0.95, RMSEA = 0.092, and two-factor, $\chi^2(151) = 1271.30$, $p < .001$, CFI = 0.98, TLI = 0.97, RMSEA = 0.073, model. The Sample 2 results were nearly identical based on the eigenvalues (first three eigenvalues were 13.91, 1.20, and 0.95) and the model fit statistics for the one-factor, $\chi^2(170) = 1282.55$, $p < .001$, CFI = 0.95, TLI = 0.95, RMSEA = 0.096, and two-factor, $\chi^2(151) = 793.45$, $p < .001$, CFI = 0.97, TLI = 0.96, RMSEA = 0.077, model. Based on these results, there is no evidence of more than two factors in the data.

For comparison purposes with previous PCL–5 research using MLR estimation and current WLSMV results, and because M*plus* does not allow for parallel analyses with WLSMV estimation, EFAs were also conducted assuming continuous variables with MLR estimation. The Sample 1 results[6] also revealed a two-factor MLR solution based on the estimated (and simulated) eigenvalues of 10.81 (1.22), 1.44 (1.18), and 1.13 (1.15) from the parallel analysis. However, the Sample 1 MLR EFA estimation model did not fit the data well for the one-factor, $\chi^2(170) = 1538.50$, $p < .001$, CFI = 0.81, TLI = 0.79, RMSEA = 0.076, or two-factor, $\chi^2(151) = 890.19$, $p < .001$, CFI = 0.90, TLI = 0.87, RMSEA = 0.059, model. Interestingly, the MLR EFA model did not fit the data well until a five-factor model was estimated, $\chi^2(100) = 313.06$, $p < .001$, CFI = 0.97, TLI = 0.94, RMSEA = 0.039, which clearly contradicts the eigenvalues, parallel analysis, and WLSMV results.

From these results, it is clear that the estimation method has a significant impact on the conclusions drawn when only focusing on the global model fit statistics, as the eigenvalues and parallel analysis told a more consistent story across the estimation methods. Although this analysis might leave researchers perplexed, it is important to rely on statistical theory. First, MLR can perform poorly with skewed ordered categorical data, and second, facets might exist in the data creating model misspecification. Based on the current data and previous factor analysis research comparing MLR and WLSMV, one could make a strong case for the use of WLSMV as a more appropriate estimation method with PCL–5 data.

The factor loading results based on the one-factor WLSMV EFA models are provided in Table 3 for Samples 1 and 2, with the results showing that each item loads on the first factor (see one-factor model results). Although there is some evidence for a two-factor model, the factors are highly correlated (see Table 3) and some of the items have larger ($\lambda > .30$) cross-loadings (e.g., Items 9 & 11). Several other items (5, 8, 16, & 18) also present some concern, with cross-loadings larger than desired (i.e., $.20 < \lambda < .30$). Collectively, these results demonstrate poor discriminant validity as the interfactor correlation suggests factors to be nearly indistinguishable and several cross-loadings imply items that measure both factors (although to varying degrees).

Overall, the dimensionality analyses and EFA results provide greater evidence for a one-factor model. Statistically, the eigenvalues and parallel analyses indicated a very weak second factor (perhaps attributable to only a few items or common method variance) and, psychometrically, a two-factor model lacks discriminant validity and would likely not lead to improved generalizability or predictive validity.

Bayesian CFA

Researchers (Hayduk, 2014a, 2014b) have argued that any model misfit (i.e., having a statistically significant χ^2) is a concern. To look deeper into the misspecification of the one-factor model and avoid concerns associated with modification indexes, a BCFA model can prove useful, as it produces results

[6]The Sample 2 results are nearly identical and reached the same conclusions and, therefore, are not presented here. These results are available from the corresponding author.

that better reflect substantive theories (Asparouhov et al., 2015; Muthén & Asparouhov, 2012) and provides a greater understanding of where model misfit occurs. In an effort to better understand any model misfit within a one-factor model (recall, for a one-factor model the only misfit can occur with the correlated residual variances or residual covariances), a BCFA model was conducted using Sample 1.

Model estimation. Using Bayesian estimation with a polychoric correlation matrix, the BCFA model was estimated using normal priors for the residual covariances [N(0, .05), thus the 95% CI was ± 0.44] and program defaults (i.e., noninformative priors) for the others. Larger prior covariances were selected to allow greater model flexibility and to avoid placing too much emphasis on the prior distribution rather than the data. After fixing the factor mean and variance at 0 and 1, respectively, to standardize the prior distributions and set the latent factor scale, all factor loadings, residual variances, and residual covariances were estimated. A well-fitting model is expected to be statistically nonsignificant, with Muthén and Asparouhov (2012) indicating values greater than .10, .05, or .01 are acceptable for most applications.

BCFA results. The BCFA model produced a good model fit; 95% CI for $\Delta\chi^2$ was [−52.96, to 66.77], posterior predictive p-value (PPP) = 0.39. As expected, the standardized factor loadings were all large ($\lambda \geq .69$) and statistically significant ($p < .001$). Several residual covariances were also large ($\delta > .30$) and produced the presence of a second factor. These large residual covariances were as follows: Item 1 with Item 2, $\delta = .48$; Item 6 with Item 7, $\delta = .56$; Item 10 with Item 11, $\delta = .38$; Item 3 with Item 13, $\delta = -.31$; Item 12 with Item 13, $\delta = .60$; Item 12 with Item 14, $\delta = .45$; Item 13 with Item 14, $\delta = .54$; Item 17 with Item 18, $\delta = .52$; Item 19 with Item 20, $\delta = .45$. Interestingly, items with higher residual covariances tended to be neighboring items, thus suggesting a concern for common method variance. In this case, people might be responding to items based on how they responded to the previous item.

These findings provide significant insight that might have been missed using the modification indexes, as many researchers would likely stop estimating parameters once the AFIs reached acceptable levels, thus failing to see this pattern in the results. These residual covariance pairs imply that Items 12 through 20 are either an artifact of common method variance or provide a very weak secondary factor (also evident by the rather small second eigenvalue). Given the large standardized factor loading sizes, it is very likely that this "second factor" does not represent a construct of any practical value. This follows the suggestion of Hayduk (2014b), who indicated that researchers need to be careful about overfactoring their data and creating factors that do not represent unique constructs.

Table 4. Number of model misspecifications at the factor loading, correlated residual, and overall (factor loadings plus correlated residuals) level using Sample 1.

	δ of .20			δ of .40		
	Factor loadings	Residual covariance	Overall	Factor loadings	Residual covariance	Overall
CFA one-factor	0	2	2	0	0	0
CFA two-factor	8	1	9	1	0	1
CFA four-factor *DSM–5*	25	1	26	15	0	15
CFA four-factor dysphoria	29	1	30	14	0	14
CFA five-factor dysphoria arousal	29	0	29	17	0	17
CFA six-factor externalizing behaviors	35	0	35	19	0	19
CFA six-factor anhedonia	30	0	30	10	0	10
CFA seven-factor hybrid	38	0	38	14	0	14

Note. CFA = confirmatory factor analysis. The number of model misspecifications is based on the following criteria: a misspecification (δ) of at least .20 or .40, a Type I error of .05, and power equal to .80.

Model misspecification

To complement the preceding results, model misspecification was also examined with the Saris–Satorra–van der Veld method. As discussed by Saris et al. (2009), global fit indexes do not indicate specific sources of model misspecification. Model fit is influenced greatly by incidental factors (e.g., sample size) that are unrelated with the model misspecification (Saris et al., 2009). An alternative to the goodness-of-fit test is to turn attention to investigating whether specific misspecifications are present in the model. They further stated, "According to our definition, a model that contains one or more relevant misspecifications is not a good model" (p. 570). To implement the Saris–Satorra–van der Veld procedure, Jrule was used to examine the power and significance of potential cross-loadings by setting the misspecification cutoff, δ, to .20 and .40, Type 1 error rate to .05, and power to .80 (see Saris et al., 2009) using Sample 1.

The PCL–5 models from Table 2 were tested for model misspecification at the factor loading and residual covariance levels, which are actually the only two areas in which misspecification could occur with these CFA models. Table 4 results indicate a clear relationship between the number of factors estimated and the number of model misspecifications, with Pearson correlations of .95 and .79 using the δ of .20 and .40, respectively. These results imply that the one-factor model is likely best, as the model could be significantly improved with only zero or two modifications depending on the size of δ. Moreover, this analysis suggests the second dimension (corresponding to the second eigenvalue) is likely an artifact of a few item pairs sharing common variance unrelated to the factor, which is essentially a form of dimensionality. In fact, after correlating two residual variances (Item 6 with Item 7 and Item 17 with Item 18), the one-factor model fit the data well, $\chi^2(102) = 1843.39$, $p < .001$, CFI = 0.96, TLI = 0.95, RMSEA = 0.084, and was a significant improvement over the one-factor model without these residual covariances, $\Delta\chi^2(2) = 239.34$, $p < .001$. Collectively, these results also provide evidence that a one-factor model is likely superior.

Bifactor models

There are numerous indexes for testing factor model reliability, dimensionality, and stability with bifactor models (see Reise, Scheines, Widaman, & Haviland, 2013; Rodriguez et al., 2016). In regard to omega (ω) reliability, these statistics include the reliability coefficients for the general/hierarchical (H) factor (ω, considers all item loadings or reliability of the general and specific factors) and specific (S) factors (ω_S, which considers only item loadings on that respective factor or the reliability of the specific factors) that represent the composite scores associated

Table 5. Bifactor model indexes for the two-factor and four-factor *DSM–5* bifactor model.

	ECV_1	ECV_2	ω/ω_S	ω_H/ω_{HS}	Relative ω	H	FD
Two-factor bifactor model							
General factor	.84	.84	.99	.92	.94	.98	.99
Specific Factor 1	.12	.23	.98	.22	.22	.71	.93
Specific Factor 2	.04	.08	.97	.00	.00	.40	.89
Four-factor *DSM–5* bifactor model							
General factor	.87	.87	.99	.96	.97	.98	.99
Specific R	.03	.12	.96	.10	.11	.38	.82
Specific AA	.02	.11	.93	.08	.09	.24	.82
Specific NACM	.03	.09	.96	.02	.02	.37	.88
Specific AR	.05	.18	.94	.16	.17	.49	.83

Note. ECV = explained common variance; H = construct replicability; FD = factor determinacy; R = re-experiencing cluster; AA = avoidance cluster; NACM = negative alterations in cognitions and mood; AT = alternations in arousal and activity. Both ECV_1 and ECV_2 are measures of the strength of a specific factor relative to all explained variance of all items; however, ECV_1 also includes those loadings not on the specific factor of interest (i.e., rather than only the explained variance of the items loading on that specific factor, as in ECV_2).

with multiple common factors. However, reliability measures (ω_H and ω_{HS}) are also available that only reflect variance related to a single factor. The ω_H and ω_{HS} assess the general factor (i.e., only the reliability of the general factor) and specific factors (i.e., only the reliability of the specific factors), respectively, after adjusting and partitioning out the general factor. An associated measure is the relative ω, which measures the ratio of these statistics (i.e., ω_H divided by ω for the general factor and ω_{HS} divided by ω for to the specific factors).

The explained common variance (ECV) measures the strength of that factor (whether general or specific) relative to all the explained item variance, with the item-level explained common variance (IECV) measuring the strength (i.e., factor loading magnitudes) of the bifactor loadings on the general factor relative to the specific factor loadings (thus, high IECV values provide initial unidimensionality). Also presented here are the factor determinacy (FD, which is the correlation between factor scores and the factors), construct replicability (H, which is construct replicability coefficient proposed by Hancock and Mueller, (2001), and percentage of uncontaminated correlations (PUC, which measures the percentage of covariance terms that only reflect variance from the general factor).

Although the standards for these bifactor indexes are not universally agreed on and tend to interact with each other (Reise, Scheines, et al., 2013), what follows are some commonly used standards. ECV values greater than .85 on the general

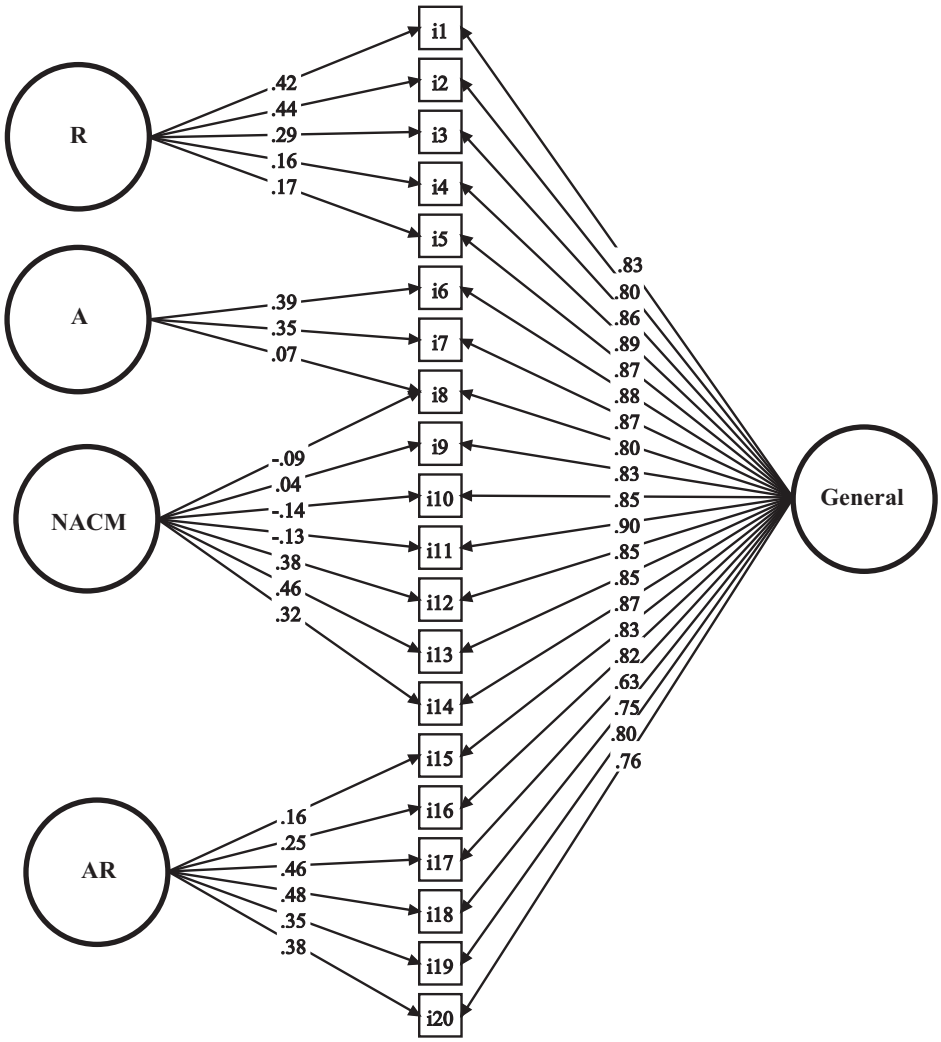

Figure 4. The factor model and standardized factor loadings for the four-factor *DSM–5* bifactor model, which includes specific factors of R (re-experiencing cluster), A (avoidance cluster), NACM (negative alterations in cognitions and mood), and AR (alterations in arousal and reactivity).

factor (and smaller specific factor ECVs) suggest the measure is sufficiently unidimensional; however, this assumes the PUC is less than .80. IECV values closer to one imply that the item reflects more of the general dimension, with a value of .50 often being the lower bound criteria. If most of the items have values above .80, this generally suggests that a unidimensional measure is more fitting.

When the ω_H is large (e.g., > .80) and the ω_{HS} are small (e.g., < .50) evidence is provided that the general factor is more reliable than specific factors. For example, when the ω_H is high, it suggests that the general factor is reliable and can be used in practice, whereas a small ω_{HS} would indicate that the specific factor is unreliable and cannot be used in practice. To our knowledge, no guidelines exist for the relative ω. It is also worth noting that when the average relative bias (ARB or the difference between the general and unidimensional factor loadings) is small, this also provides evidence of unidimensionality. PUC is a measure of unidimensionality, with PUCs greater than .70 often providing evidence of unidimensionality. FD values less than .90 and H values greater than .80 are preferable.

Using the full sample (N = 1,403) and WLSMV estimation, two bifactor models were estimated to determine whether a general or unidimensional factor was more appropriate than either the traditional two-factor or four-factor *DSM–5* model in Table 2. Although bifactor models could be estimates for the remaining models in Table 2, it was determined that these results would be repetitive and not provide any new information related to dimensionality. Results for the two-factor and four-factor *DSM–5* model were selected because they match the number of factors best supported by our dimensionality analyses and best correspond to theory, respectively. Note that the use of a bifactor model was further supported by a first to second eigenvalue ratio of 14.10 to 1.20, as this large of a ratio between the first and second eigenvalues provides additional evidence of a bifactor model (see Reise et al., 2010; Reise, Scheines, et al., 2013).

Two-factor bifactor model

The two-factor bifactor model (see Figure 3) was first tested to determine whether the bifactor model is more statistically appropriate than the traditional two-factor model (see Table 2). This two-factor bifactor model fit the data well, $\chi^2(150) = 944.94$, $p < .001$, CFI = 0.98, TLI = 0.98, RMSEA = 0.06, with the standardized factor loadings on the specific factors being considerably smaller than the general factor. However, this difference was more noticeable for Factor 2. These factor loading results suggest that although the items correlate consistently well with the general model, this is not always true for the specific factors.

To provide a more complete evaluation of the bifactor model, various bifactor indexes were examined using the previously outlined standards (see earlier). The ECV (0.84; see Table 5) and ARB (0.078) exceeded the minimum criteria (ECV = 0.70 and ARB = 0.10; Stucky & Edelen, 2014) to meet the definition of a bifactor model, with the specific factor ECVs being extremely small (see Table 5). In addition, the average IECV was .85 (SD = 0.11, minimum = 0.65, maximum = 1.00), with 75% of the IECV being greater than .80. This also provides support for the bifactor model. The PUC for this model was .53.

The ω_H and ω_{HS} indicated that the reliability was high for the general factor (ω_H = 0.99), but low for the specific factors ω_{SH} (see Table 5). These results imply that although the general factor is reliable, the specific factors are not after adjusting for the general factor. Although of less interest here, it is worth noting that the FD statistic exceeded .90 (see Gorsuch, 1983; Rodriguez, Reise, & Haviland, 2016) for the general and specific factors, but the H statistic was only acceptable (i.e., > .80) for the general factor and not the specific factors (see Table 5). Collectively, these results suggest that a bifactor model is more statistically appropriate than a model with two specific factors and research should consider using the total PCL–5 score.

Four-factor DSM–5 bifactor model

This bifactor model also fit the data well, $\chi^2(149) = 1233.63$, $p < .001$, CFI = 0.97, TLI = 0.97, RMSEA = 0.08, and resulted in much lower standardized factor loading on the specific factors significantly than the bifactor (see Figure 4). The ECV (.87, see Table 5) and ARB (.03) also exceeded the minimum criteria (ECV = .70 and ARB = .10) for evidence of a bifactor model and the specific factor ECVs were much smaller (see Table 5). The average IECV was 0.87 (SD = 0.10, minimum = 0.66, maximum = 1.00), with 80% of the IECV being greater than .80. The ω_H for the general factor was once again large, with much smaller ω_{HS} for the specific factors (see Table 5). The PUC for this model was .77. The FD statistic exceeded .90 for the general, but not the specific factors. The H statistic was also only acceptable for the general factor and not the specific factors (see Table 5). As with the previous bifactor results, these analyses also indicate that a bifactor model is more statistically appropriate than the four-factor *DSM–5* model.

Conclusions

The intent of this article and study was to encourage researchers to employ psychometrically and statistically rigorous methods to retain the "optimal" number of factors and estimate an appropriate factor structure. To begin truly dissecting clinically relevant latent factor structures, researchers must judiciously diagnose model fit, provide potential explanations to differences in results within and across studies, and take the next steps in model exploration.

As Cattell (1966) stressed, searching for the "correct" number of factors is an exercise in futility, and only detracts from a thorough investigation of factor structures that are worthwhile to retain for the optimal number of factors and explicitly stated scientific goal (Cudeck & Henly, 2003; Preacher et al., 2013). For this reason, researchers should follow the recommendations of Preacher and Merkle (2012) to "find a useful approximating model that (a) fits well, (b) has easily interpretable parameters, (c) approximates reality in as parsimonious a fashion as possible, and (d) can be used as a basis for inference and prediction" (p. 1). Our findings challenge previous PCL–5 research that runs contrary to the recommendations of Preacher and Merkle (2012), as we found little statistical and psychometric evidence for more than two factors. In fact, a stronger argument exists for a single-factor model based on the

parallel analyses, model fit statistics, lack of discriminant validity, model misspecification analyses, BCFA, and bifactor results. Moreover, several of the previously proposed models are significantly flawed from a psychometric (e.g., fewer than three items per factor) and statistical (e.g., inappropriate model estimation) perspective. The resultant danger of overreliance on simple global fit statistics is that excessively complex models tend to overfit the data; thus, using global goodness-of-fit indexes as the only model selection criteria is analogous to "*p*-hacking" and is not prudent (Browne & Cudeck, 1992; Head, Holman, Lanfear, Kahn, & Jennions, 2015; Roberts & Pashler, 2000; Siegfried, 2010; Wasserstein & Lazar, 2016).

If complex models result in more accurate cross-validated and better prediction, then it makes sense to use the complex models, but if they do not, then simpler models should be used. Further, Ropovik (2015) reviewed 11 psychological journals and found that researchers have a propensity to accept theoretically complex models. We concur with this finding and found that PCL–5 researchers tend to overextract factors that are neither statistically nor psychometrically justified and that seem to lean heavily toward verisimilitude. Further, depending on the diagnostic scoring algorithm (e.g., number of symptoms required for each model-derived factor), different psychometrically and statistically derived models can also have an impact on diagnostic criteria (e.g., diagnostic algorithms based on different models), which in turn affects PTSD prevalence rates (see Murphy et al., 2017; Shevlin, Hyland, Karatzias, Bisson, & Roberts, 2017; Wortmann et al., 2016). Thus, thoroughly searching for the "correct" or "best" factor analytic model is a statistically and clinically important endeavor.

In conclusion, the intent of this article is to encourage researchers seeking to evaluate the psychometric properties of measures to take a more rigorous and holistic approach to measurement evaluation. Our hope also is that researchers carefully consider the model estimation method, the practical utility of the factors (e.g., how well it predicts the desired outcomes), the intended purpose of the measure (i.e., generalizability vs. verisimilitude), and justify their measurement and modeling decisions based on good statistical and psychometric practices. Without these steps, researchers are in danger of flooding research journals with inconsistent and potentially incorrect models, which will only slow the progress of science and hurt the people these models and measures were designed to help.

Acknowledgments

The authors would like to give special thanks to Tanya Goodman for assistance with measure administration, data collection, and institutional review board approval. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the U.S. Air Force, the Department of Defense, or the U.S. Government.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*, 3–25. doi:10.1037/amp0000191

Armour, C., Műllerová, J., & Elhai, J. D. (2016). A systematic literature review of PTSD's latent structure in the *Diagnostic and Statistical Manual of Mental Disorders*: DSM–IV to DSM–5. *Clinical Psychology Review*, *44*, 60–74. doi:10.1016/j.cpr.2015.12.003

Armour, C., Tsai, J., Durham, T. A., Charak, R., Biehn, T. L., Elhai, J. D., & Pietrzak, R. H. (2015). Dimensional structure of DSM–5 posttraumatic stress symptoms: Support for a hybrid anhedonia and externalizing behaviors model. *Journal of Psychiatric Research*, *61*, 106–113. doi:10.1016/j.jpsychires.2014.10.012

Ashbaugh, A. R., Houle-Johnson, S., Herbert, C., El-Hage, W., & Brunet, A. (2016). Psychometric validation of the English and French versions of the Posttraumatic Stress Disorder Checklist for DSM–5 (PCL–5). *PLoS ONE*, *11*, e016164. doi:10.1371/journal.pone.0161645

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438. doi:10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data* (Technical Report). Retrieved from http://www.statmodel.com/download/GstrucMissingRevision.pdf

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances. *Journal of Management*, *41*, 1561–1577. doi:10.1177/0149206315591075

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. doi:10.1038/533452a

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *3*, 186–203. doi:10.1207/s15328007sem1302_2

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, *105*, 467–477. doi:10.1037/0033-2909.105.3.467

Biehn, T. L., Elhai, J. D., Seligman, L. D., Tamburrino, M., Armour, C., & Forbes, D. (2013). Underlying dimensions of DSM–5 posttraumatic stress disorder and major depressive disorder symptoms. *Psychological Injury and Law*, *6*, 290–298. doi:10.1007/s12207-013-9177-4

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The Posttraumatic Stress Disorder Checklist for DSM–V (PCL–5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, *28*, 489–498. doi:10.1002/jts.22059

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

Bovin, M. J., Marx, B. P., Weathers, F. W., Gallagher, M. W., Rodriguez, P., Schnurr, P. P., & Keane, T. M. (2016). Psychometric properties of the PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders–Fifth Edition (PCL–5) in veterans. *Psychological Assessment*, *28*, 1379–1391. doi:10.1037/pas0000254

Bozdogan, H. (2000). Akaike's information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology*, *44*, 62–91. doi:10.1006/jmps.1999.1277

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. doi:10.1207/S15327906MBR3601_05

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230–258. doi:10.1177/0049124192021002005

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for multidimensionality and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 247–271). Gottingen, Germany: Hogrefe.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276. doi:10.1207/s15327906mbr0102_10

Chappelle, W., Goodman, T., Reardon, L., & Thompson, W. (2014). An analysis of post-traumatic stress symptoms in United States Air Force drone operators. *Journal of Anxiety Disorders*, *28*, 480–487. doi:10.1016/j.janxdis.2014.05.003

Coenders, G., Satorra, A., & Saris, W. E. (1997). Alternative approaches to structural modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling*, *4*, 261–282. doi:10.1080/10705519709540077

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*, 147–168. doi:10.1177/1094428103251541

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519. doi:10.1037/0033-2909.109.3.512

Cudeck, R., & Henly, S. J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods*, *8*, 378–383. doi:10.1037/1082-989X.8.3.378

Davidson, J. R. T., & Foa, E. B. (1991). Diagnostic issues in posttraumatic stress disorder: Considerations for the DSM–IV. *Journal of Abnormal Psychology*, *100*, 346–355. doi:10.1037/0021-843X.100.3.346

Eddinger, J. R., & McDevitt-Murphy, M. E. (2017). A confirmatory factor analysis of the PTSD Checklist 5 in veteran and college student samples. *Psychiatry Research*, *255*, 219–224. doi:10.1016/j.psychres.2017.05.035

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research*, *63*, 324–327. doi:10.1016/j.jbusres.2009.05.003

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

Frewen, P. A., Brown, M. F. D., Steuwe, C., & Lanius, R. A. (2015). Latent profile analysis and principal axis factoring of the DSM–5 dissociative subtype. *European Journal of Psychotraumatology*, *6*, 26406. doi:10.3402/ejpt.v6.26406

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling*, *3*, 62–72. doi:10.1080/10705519609540030

Geweke, J. F., & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, *75*, 133–137. doi:10.1080/01621459.1980.10477442

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*, 532–560. doi:10.1207/s15327752jpa6803_5

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 143–164). Hoboken, NJ: Wiley.

Gray-Little, B., & Kaplan, D. A. (1998). Interpretation of psychological tests in clinical and forensic evaluations. In J. H. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.), *Test interpretation and diversity: Achieving equity in assessment* (pp. 141–178). Washington, DC.: American Psychological Association.

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

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Hayashi, K., Bentler, P. M., & Yuan, K. H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling*, *14*, 505–526. doi:10.1080/10705510701301891

Hayduk, L. A. (2014a). Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement*, *74*, 905–926. doi:10.1177/0013164414527449

Hayduk, L. A. (2014b). Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology*, *14*, 124. doi:10.1186/1471-2288-14-124

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*, e1002106. doi:10.1371/journal.pbio.1002106

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling*, *14*, 361–390. doi:10.1080/10705510701301602.

Hoge, C. W., Riviere, L. A., Wilk, J. E., Herrell, R. K., & Weathers, F. W. (2014). The prevalence of post-traumatic stress disorder (PTSD) in US combat soldiers: A head-to-head comparison of DSM–5 versus DSM–IV–TR symptom criteria with the PTSD Checklist. *The Lancet Psychiatry*, *1*, 269–277. doi:10.1016/S2215-0366(14)70235-4

Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453. doi:10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Kaplan, R. M., & Saccuzzo, D. P. (2008). *Psychological testing principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth.

Katz, R. W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics*, *23*, 243–249. doi:10.2307/1267787

Keane, T. M., Rubin, A., Lachowicz, M., Brief, D., Enggasser, J. L., Roy, M., … Rosenbloom, D. (2014). Temporal stability of DSM–5 posttraumatic stress disorder criteria in a problem-drinking sample. *Psychological Assessment*, *26*, 1138–1145. doi:10.1037/a0037133

King, R. V., North, C. S., Surís, A., & Smith, R. P. (2016). The evolution of PTSD criteria across editions of DSM. *Annals of Clinical Psychiatry: Official Journal of the American Academy of Clinical Psychiatrists*, *28*, 197–208.

Konecky, B., Meyer, E. C., Kimbrel, N. A., & Morissette, S. B. (2016). The structure of DSM–5 posttraumatic stress disorder symptoms in war veterans. *Anxiety, Stress, & Coping*, *29*, 497–506. doi:10.1080/10615806.2015.1081178

Krüger-Gottschalk, A., Knaevelsrud, C., Rau, H., Dyer, A., Schäfer, I., Schellong, J., & Ehring, T. (2017). The German version of the Posttraumatic Stress Disorder Checklist for DSM–5 (PCL–5): Psychometric properties and diagnostic utility. *BMC Psychiatry*, *17*, 1–9. doi:10.1186/s12888-017-1541-6

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949. doi:10.3758/s13428-015-0619-7

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education*, *2*, 17–38. doi:10.1504/IJQRE.2014.060972

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods*, *4*, 192–221. doi:10.1037/1082-989X.4.2.192

Liu, P., Wang, L., Cao, C., Wang, R., Zhang, J., Zhang, B., … Elhai, J. D. (2014). The underlying dimensions of DSM–5 posttraumatic stress disorder symptoms in an epidemiological sample of Chinese earthquake survivors. *Journal of Anxiety Disorders*, *28*, 345–351. doi:10.1016/j.janxdis.2014.03.008

MacCallum, R. C., Rosnowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. doi:10.1037/0033-2909.111.3.490

Marsh, H. W., Liem, G. A., Martin, A. J., Morin, A. J., & Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, *29*, 322–346. doi:10.1177/0734282911406657

Marsh, H. W., Muthén, B., Asparouhov, A., Lüdtke, O., Robitzsch, A., Morin, A. J. S., … Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439–476. doi:10.1080/10705510903008220

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, *27*, 82–99. doi:10.1111/j.2044-8317.1974.tb00530.x

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*, 247–255. doi:10.1037/0033-2909.107.2.247

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Hillsdale, NJ: Erlbaum.

McSweeney, L. B., Koch, E. I., Saules, K. K., & Jefferson, S. (2016). Exploratory factor analysis of *Diagnostic and Statistical Manual*, 5th edition, criteria for posttraumatic stress disorder. *Journal of Nervous and Mental Disease*, *204*, 9–14. doi:10.1097/NMD.0000000000000390

Mordeno, I. G., Nalipay, M. J. N., Sy, D. J. S., & Luzano, J. G. C. (2016). PTSD factor structure and relationship with self-construal among internally displaced persons. *Journal of Anxiety Disorders*, *44*, 102–110. doi:10.1016/j.janxdis.2016.10.013

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York, NY: McGraw-Hill.

Mulaik, S. A. (2001). The curve-fitting problem: An objectivist view. *Philosophy of Science*, *68*, 218–241. doi:10.1086/392874

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman & Hall/CRC.

Murphy, S., Hansen, M., Elklit, A., Yong Chen, Y., Raudzah Ghazali, S., & Shevlin, M. (2017). Alternative models of DSM–5 PTSD: Examining diagnostic implications. *Psychiatry Research*. Advance online publication. doi:10.1016/j.psychres.2017.09.011

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. doi:10.1037/a0026802

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170–11175. doi:10.1073/pnas.170283897

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In J. Wixted & H. Pashler (Eds.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (3rd ed., Vol. 4, pp. 429–459). New York, NY: Wiley.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi:10.1126/science.aac4716

Padilla, A. M., & Borsato, G. N. (2008). Issues in culturally appropriate psychoeducational assessment. In L. Suzuki & J. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological and educational assessment* (3rd ed., pp. 5–21). San Francisco, CA: Jossey-Bass.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of nontrivial axes revisited. *Computational Statistics & Data Analysis*, *49*, 974–997. doi:10.1016/j.csda.2004.06.015

Pietrzak, R. H., Tsai, J., Armour, C., Mota, N., Harpaz-Rotem, I., & Southwick, S. M. (2015). Functional significance of a novel 7-factor model of DSM–5 PTSD symptoms: Results from the National Health and Resilience in Veterans study. *Journal of Affective Disorders*, *174*, 522–526. doi:10.1016/j.jad.2014.12.007

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491. doi:10.1037/0033-295X.109.3.472

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*, 227–259. doi:10.1207/s15327906mbr4103_1

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, *17*, 1–14. doi:10.1037/a0026804

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, *48*, 28–56. doi:10.1080/00273171.2012.710386

Raubenheimer, J. (2004). An item selection procedure to maximise scale reliability and validity. *SA Journal of Industrial Psychology*, *30*. doi:10.4102/sajip.v30i4.168

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. doi:10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*, 129–140. doi:10.1080/00223891.2012.725437

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. doi:10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5–26. doi:10.1177/0013164412449831

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*, 287–297. doi:10.1037/1040-3590.12.3.287

Revelle, W. (n.d.). *An introduction to psychometric theory with applications in R*. Manuscript in preparation.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367. doi:10.1037/0033-295X.107.2.358

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*, 223–237. doi:10.1080/00223891.2015.1089249

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology*, *6*, 1715. doi:10.3389/fpsyg.2015.01715

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561–582. doi:10.1080/10705510903203433

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*, 73–103. doi:10.1080/00273170903504810

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

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*, 5–30. doi:10.1177/0265532208097335

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*, 304–321. doi:10.1177/0734282911406653

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, *71*, 95–113. doi:10.1177/0013164410387348

Shevlin, M., Hyland, P., Karatzias, T., Bisson, J. I., & Roberts, N. P. (2017). Examining the disconnect between psychometric models and clinical reality of posttraumatic stress disorder. *Journal of Anxiety Disorders*, *47*, 54–59. doi:10.1016/j.janxdis.2017.02.006

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, *63*, 117–126. doi:10.1093/biomet/63.1.117

Siegfried, T. (2010). Odds are, it's wrong: Science fails to face the shortcomings of statistics. *Science News*, *177*, 26–37. doi:10.1002/scin.5591770721

Southwick, S. M., Yehuda, R., & Giller, E. (1993). Personality disorders in treatment seeking Vietnam combat veterans with post-traumatic stress disorder. *American Journal of Psychiatry*, *150*, 1020–1023. doi:10.1176/ajp.150.7.1020

Stucky, B. D., & Edelen, M. O. (2014). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183–206). London, UK: Taylor & Francis.

Taylor, M. A., & Pastor, D. A. (2007). A confirmatory factor analysis of the student adaptation to college questionnaire. *Educational & Psychological Measurement*, *67*, 1002–1018. doi:10.1177/0013164406299125

Tsai, J., Harpaz-Rotem, I., Armour, C., Southwick, S. M., Krystal, J. H., & Pietrzak, R. H. (2015). Dimensional structure of DSM–5 posttraumatic stress disorder symptoms: Results from the National Health and Resilience in Veterans study. *Journal of Clinical Psychiatry*, *76*, 546–553. doi:10.4088/JCP.14m09091

van der Eijk, C., & Rose, J. (2015). Risky business: Factor analysis of survey data—Assessing the probability of incorrect dimensionalisation. *PLoS ONE*, *10*, e0118900. doi:10.1371/journal.pone.0118900

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231–251. doi:10.1037/1082-989X.3.2.231

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133. doi:10.1080/00031305.2016.1154108

Weathers, F., Litz, B., Herman, D., Huska, J., & Keane, T. (1993). *The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility*. Scale available from the National Center for PTSD at http://www.ptsd.va.gov.

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). *The PTSD Checklist for DSM–5 (PCL–5)*. Retrieved from https://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58–79. doi:10.1037/1082-989X.12.1.58

Wortmann, J. H., Jordan, A. H., Weathers, F. W., Resick, P. A., Dondanville, K. A., Hall-Clark, B., … Litz, B. T. (2016). Psychometric analysis of the PTSD Checklist–5 (PCL–5) among treatment-seeking military service members. *Psychological Assessment*, *28*, 1392–1403. doi:10.1037/pas0000260

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*, 377–392. doi:10.1177/0734282911406668

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*, *47*, 756–772. doi:10.3758/s13428-014-0499-2

Young, G. (2016). PTSD in Court I: Introducing PTSD for court. *International Journal of Law and Psychiatry*, *49*, 238–258. doi:10.1016/j.ijlp.2016.10.012