ijbd

International Journal of Behavioral Development
2021, Vol. 45(1) 40–50

DOI: 10.1177/0165025419896620
journals.sagepub.com/home/jbd
SAGE

Psychometric models for scoring multiple reporter assessments: Applications to integrative data analysis in prevention science and beyond

Patrick J. Curran,[1] A. R. Georgeson,[1] Daniel J. Bauer,[1] and Andrea M. Hussong[1]

Abstract

Conducting valid and reliable empirical research in the prevention sciences is an inherently difficult and challenging task. Chief among these is the need to obtain numerical scores of underlying theoretical constructs for use in subsequent analysis. This challenge is further exacerbated by the increasingly common need to consider multiple reporter assessments, particularly when using integrative data analysis to fit models to data that have been pooled across two or more independent samples. The current article uses both simulated and real data to examine the utility of a recently proposed psychometric model for multiple reporter data called the trifactor model (TFM) in settings that might be commonly found in prevention research. Results suggest that numerical scores obtained using the TFM are superior to more traditional methods, particularly when pooling samples that contribute different reporter perspectives.

Keywords

Integrative data analysis, moderated nonlinear factor analysis, trifactor model, psychometrics, scoring, multiple reporter assessments

As you likely know well yourself, conducting empirical research in the prevention sciences and related disciplines encompasses nearly every quantitative and methodological challenge that can be imagined. Common components include experimental and quasi-experimental designs, causal inference, trajectory analysis, multisite designs, missing data, non-normally distributed data, discretely scaled data, clustered observations, nonlinearity, mediation, moderation, and observed or latent subgroup heterogeneity, among many (many) others. These complexities are further exacerbated when using integrative data analysis (IDA) to fit models to sample data that have been pooled from two or more existing studies (Curran, 2009; Curran & Hussong, 2009; Hussong et al., 2013). Although these challenges can arise in various combinations in any given study, there remains one critically important component that is common to all empirical research applications in the prevention sciences: *measurement* and *scoring*.

Briefly, measurement is the psychometric modeling of the relation between a set of observed items and an underlying theoretical construct, and scoring is the computation of numerical indices (or *scale scores*) that are used in subsequent hypothesis testing (e.g., Bandalos, 2018). Over a century ago, Thorndike (1918) famously noted "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality." Indeed, all that we seek to obtain to enhance and expand our understanding of the course, causes, and consequences of human behavior fundamentally rests on the valid and reliable numerical assessment of the constructs under study. The extent to which any given measure fails to properly assess what it purports irreparably limits our ability to make valid inferences about the processes under study (Shadish et al., 2002).

Although a shared component of virtually all disciplines of scientific inquiry, measurement and scoring is especially salient in the empirical examination of prevention or intervention programs (e.g., Collins & Flaherty, 2006). Among a variety of challenges, an issue that is becoming increasingly relevant is the desire to obtain assessments of a target individual (say a child) from two or more independent reporters (say a parent and a teacher). Although widely used in practice, it has long been known that single reporter inextricably confounds the assessment of the underlying trait with the perspective of the reporter (Achenbach et al., 2005; Renk, 2005). For example, using the mother's report of her child's behavior by necessity limits the assessments as seen through the lens of the mother (e.g., Boyle & Pickles, 1997). Other perspectives that might provide additional insights into the child's behavior might include the father, a teacher, or a best friend or even extant data drawn from sources such as health or education records; see De Los Reyes et al. (2013) for an excellent review of these issues.

There is broad consensus on the need for multiple reporter assessments, yet the analytic methods available to incorporate these separate sources of information in some principled fashion remain limited (Achenbach, 2011; Bauer et al., 2013). Measurement and scoring of multiple reporter data becomes even more complicated

[1] University of North Carolina, USA

Corresponding author:
Patrick J. Curran, Department of Psychology and Neuroscience, University of North Carolina, Chapel Hill, NC 27599, USA.
Email: curran@unc.edu.

when pooling two or more data sources using IDA where some reporters may be present in one contributing study but not in others. The measurement and scoring of multiple reporter data drawn from multiple independent studies using an IDA framework is the focus of our work here.

Because the topic of this special issue is *prevention science*, we frame our discussion in terms of issues that arise in the empirical evaluation of prevention and intervention programs. However, all of the challenges, models, and results generalize to a broad class of experimental settings that are not directly related to studies of prevention. We begin by considering traditional methods of measurement and scoring before turning to more recent psychometric models that are well suited to the analysis of multiple reporter data.

Traditional Methods of Measurement and Scoring

Historically, there are three broad traditions that have been used to compute scale score estimates based on a set of observed items. We address each of these approaches in turn.

Naive Scores

The first is embedded in classical test theory (CTT) and is based on the premise that each observed item is an inextricable mix of true score and error. CTT has supported the use of widely used methods such as sum scores, proportion scores, and mean scores that have been a staple for empirical research for nearly a century (Bandalos, 2018). These sum scores are characterized by certain advantages, the most notable is ease in computation and interpretation. For example, for a given set of observed items (e.g., binary indicators reflecting the presence or absence of symptoms of adolescent depression), a scale score can be obtained by simply summing the items to reflect a total symptom count (or then dividing by the number of items to reflect the proportion of items endorsed).

However, these sum scores are also characterized by many disadvantages, a key one of which is the assumption that all items are measured without error and are equally related to the underlying factor. For example, endorsing an item that assesses the extent to which an adolescent feels lonely and another item that assesses an adolescent's fantasizing about taking their own life are treated as equally indicative of underlying depression. Further, there is no differential weighting of sum scores as a function of any between-person characteristics such as the child's biological sex, race, age, or treatment condition; thus all items are assumed to operate in precisely the same way for all individuals. Because traditional scores are computed in the absence of any other information, these are sometimes called *naive scores*.

Latent Variable Scores

In contrast to CTT-based scoring methods, there is a long tradition of psychometric methods focused on latent variable-based models including item response theory (IRT; e.g., Embretson & Reise, 2013) and factor analysis, both exploratory (EFA) and confirmatory (CFA) (e.g., Bandalos & Finney, 2018). Whereas naive scores treat all items as error-free and equally indicative of the underlying construct, IRT and factor analysis models estimate measurement error while also allowing items to be differentially related to the underlying latent factor (i.e., more salient items are more strongly linked to the factor and vice versa; e.g., Thissen & Orlando, 2001; Wirth & Edwards, 2007). These latent variable models can serve two purposes. First, they represent a formal psychometric expression of the underlying latent structure that gave rise to the set of observed indicators (e.g., number of factors, pattern of factor loadings, correlated residuals, etc.). Second, given a particular psychometric model, the final set of parameter estimates can be applied to the raw data to provide *factor score estimates* that are scale scores intended to reflect the underlying latent factors and are available for subsequent analysis. Unlike sum or mean scores that treat all items equally, latent variable models allow for more informative items to have a greater impact on estimated scores than do weaker ones. There is a long history of factor score estimation in the social sciences (see, Grice, 2001), yet these methods remain under used in practice (e.g., DiStefano et al., 2009).

Covariate-Informed Scores

Both the naive and latent variable approaches assume that the computation of the scores are constant across all between-person characteristics. That is, the strength of the contribution of each item to the underlying scale is the same across factors such as biological sex, race, treatment condition, age, or socioeconomic status. However, between-person characteristics might impact both the measurement and the scoring models, the omission of which may substantially bias results. In a seminal paper on this topic, Mislevy et al. (1992) argued that scores for latent variables should include the impact of relevant background variables (or *covariates*). Covariates might influence model results in two ways: in their relation to the mean and variance of the latent factor itself (called *impact*) or to the item intercept and factor loading linking each item to the factor (called differential item functioning, or DIF), or both (Holland & Wainer, 2012). Although impact and DIF can be incorporated into both the IRT and CFA through multiple group models, these are typically limited to examining the effects of membership in one of two discrete groups. However, both theory and empirical data may call for the simultaneous consideration of multiple characteristics such as biological sex, age, race, treatment condition, or their interactions. Such multivariable effects are not possible in standard two-group IRT or CFA methods (Bauer, 2017).

A comprehensive method for incorporating impact and DIF into a latent variable model using multiple covariates of varying distributional types was proposed by Bauer and colleagues (Bauer, 2017; Bauer & Hussong, 2009; Curran et al., 2014; Curran et al., 2016; Curran et al., 2018). Briefly, a set of exogenous covariates (e.g., sex, race, treatment condition) can be incorporated into the latent variable model and explicitly exert both DIF and impact effects in the estimated model as well as the resulting factor score estimates. Constraints are written for specific parameters in the CFA that shift the estimates as a function of the set of covariates (e.g., a factor loading for a particular item may vary in magnitude as a joint function of biological sex, treatment condition, and age). Simulation studies have shown these covariate-informed scores to be superior to all other scoring options (Curran et al., 2016; Curran et al., 2018). Despite these many advantages, this analytic method does not easily incorporate multiple reporter data. However, Bauer expanded this approach to address the very issue of multiple reporter assessments.

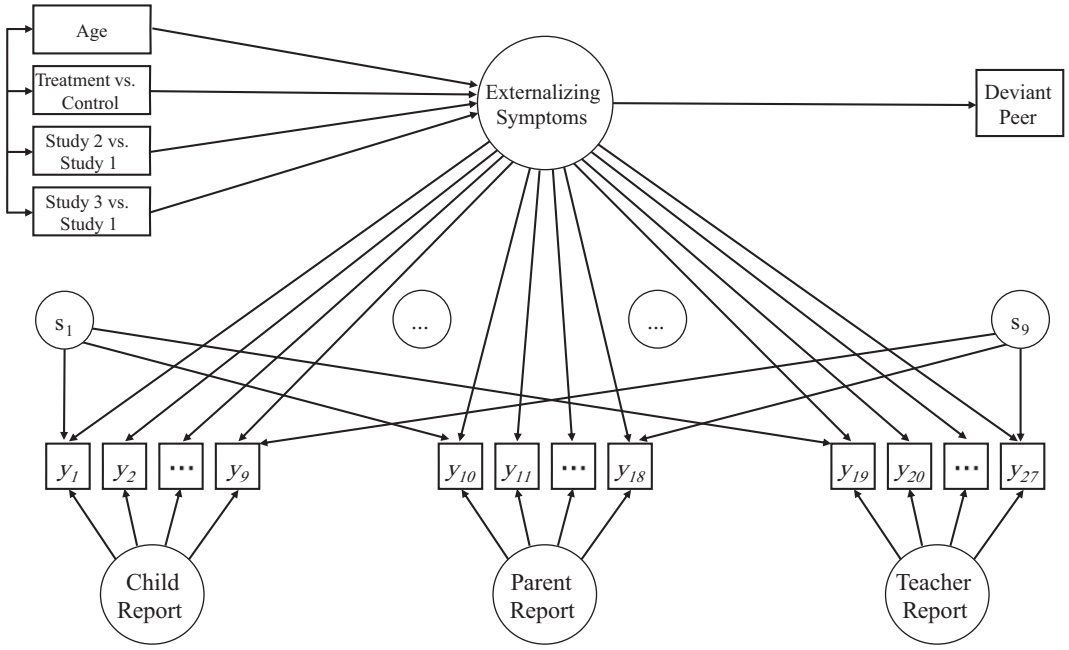

Figure 1. Trifactor Model Used for Data Generation in Monte Carlo Simulations.
Note. S = specific factor (nine total, only two shown).

The Trifactor Model

Bauer et al. (2013) proposed a novel analytic approach called the *trifactor model* (or TFM) to expand the covariate-informed factor model to explicitly include multiple reporter data. This was named as such because the model has a structure consisting of three levels of latent variables (or *factors*) that define different sources of variability contributing to a set of items. For example, say that an IDA was designed to pool sample data from prevention programs that assessed child behavior using responses to the same set of items relating to the child based on three reporters: child, parent, and teacher. The TFM defines a psychometric model to consist of three types of latent factors. The first is a *common factor* that represents the pooled information shared by all three reporters across the entire set of items; this might represent a global assessment of child depression. The second are a set of *perspective factors* that are defined for each reporter; this represents the unique perspectives of the child, parent, and teacher. Finally, the third are a set of *specific factors* that are defined for each item that is shared across the three reporters; this represents the possibility of some shared characteristic that is unique to a specific item. All factors are typically orthogonal, allowing for the separation of the unique contribution at each level. An exemplar path diagram of this model is presented in Figure 1 (described in more detail below).

The TFM can be expanded in a variety of interesting ways. First, not all reporters need to respond to the same set of items. Thus, within an IDA application, some reporters might respond to one set of items while other reporters respond to a subset of these items but respond to additional items that are unique to their own report. Second, complete-case data are not required. Thus within an IDA application some contributing samples might have three reporters available, some might have two, and some might have just one, yet these can all be combined within a single TFM (under certain assumptions about the missing data; Enders, 2010). Third, the TFM allows for the incorporation of reporter-specific characteristics as covariates in the model that help in part determine the unique reporter perspective. For example, parental alcoholism diagnosis could be included as a predictor of parent perspective and years of experience could be included as a predictor of teacher perspective. Finally, the TFM can provide factor score estimates on the common factor that represent the optimal pooled combination of all available reporters net the potentially biasing effects of the perspective and specific factors, and these scores can be used in subsequent modeling and analysis. Complete details about these and other aspects of the TFM are available in Bauer et al. (2013).

Summary

In sum, recent advances in psychometric modeling provide a principled method for computing scale scores that allow for both differential strength of relations between items and the factor and for measurement error, include the influence of exogenous covariates, and incorporate assessments drawn from multiple independent reporters. However, these psychometric models have yet to be closely examined when applied to multiple reporter data that have been pooled from two or more studies using IDA. Our goal is to examine the performance of these models using both simulated and real data under conditions that might commonly be encountered in applied research settings within the prevention sciences and related disciplines and to make empirically informed recommendations for the broader use of these methods in practice.

Simulation Study: Method

We begin by examining artificially generated data for a single large sample of data that corresponds to a known population-generating model to examine large-sample characteristics of model estimation and inference.

Artificial Data Model

Measurement model. We generated empirical data to be consistent with the conditional TFM described earlier. To enhance external validity, we designed the population-generating model to be closely representative of empirical characteristics that might commonly be encountered in applied prevention research settings, particularly when using IDA to combine data from multiple sites or samples. The core of the model was designed around the hypothetical situation in which there are three separate reporters crossed with three independent studies contributing data to the IDA. The outcome of interest is the behavior of a target child, and the three reporters are designed to represent child-, parent-, and teacher-report of child behavior. In the hypothetical scenario, each of the three reporters responds to the same nine binary items (with the child as the shared target) evaluating the presence or absence of child externalizing symptomatology (e.g., aggression, delinquency). The 9 observed items are linked to the underlying latent factor using actual item parameters drawn from measurement models fitted to real data assessing the Child Behavior Check List (Achenbach, 1999; Achenbach & Edelbrock, 1981) subscale of externalizing symptomatology.

Structural model. For the conditional TFM, the common factor representing child externalizing symptomatology was regressed on four exogenous covariates: a binary measure of treatment condition (0 = control, 1 = treatment with subjects evenly divided in each), child age (integer values from 10 to 16 and centered at age 13), and two dummy codes comparing Study 2 to Study 1 and Study 3 to Study 1. Further, the common factor predicted a continuously and normally distributed outcome that was designed to represent a measure of affiliation with delinquent peers. This model thus represents a hypothetical prevention program designed to decrease externalizing symptomatology that in turn decreases affiliation with deviant peers. See Figure 1 for a graphical representation of the final model.

Data generation. First, continuous and normally distributed "true" factor scores were drawn for all latent variables in the TFM. The scores for the *common factor* were drawn from a conditional distribution defined by the joint contribution of the four covariates (treatment, age, and the two dummy codes); covariate effects were defined as having a moderate effect size (squared semi-partial correlations ranging from .04 to .11) with a joint multiple-R^2 value of .22. The scores for the three *perspective factors* were drawn without the influence of any covariates; the child perspective was drawn from a standard normal distribution, and the parent and teacher perspective factors were drawn from a distribution with a larger mean and variance relative to the child (to reflect differential predictability across reporters). Finally, all *item specific factors* were drawn from a standard normal distribution that did not vary over reporter or study. For all factors, 5,000 true factor scores were generated within each of three studies for a total sample size of 15,000 simulated cases.[1]

Next, a logit link function was used to create the binary item responses for each of the 9 items associated with each of the three reporters drawn from each of the three studies based on the item parameters and the continuous and normally distributed true scores described earlier (see, e.g., Curran et al., 2018, Equations 3, 4, and 5). This step resulted in a single data file consisting of triplets of nine reporter-specific binary items plus the four covariates and the distal outcome.

Finally, three experimental "settings" were created. For Setting 1, we assumed that all three studies contributed all three perspectives; this can be considered the complete data condition in which all reporters are present in all studies. For Setting 2, all three perspectives were present in Study 1, but Study 2 only contributed child report and Study 3 only contributed parent report; this allowed for the examination of the potential added value of including studies in the IDA that provided only a single perspective. Finally, for Setting 3, Study 1 contributed child and parent report, Study 2 contributed child and teacher report, and Study 3 contributed parent and teacher report; this allowed for the examination of an IDA in which no study provided all three perspectives, but all three perspectives were jointly represented by the pairing of reporters in each contributing study.

Scoring models. Following the above procedures, we generated three sample data files, one for each of the three settings. We obtained numerical score estimates for child externalizing symptomatology using two different scoring models. First, we computed traditional *proportion scores* for all available reporters within each setting by computing the mean of the 9 binary items to which each reporter responded; we then computed the mean of the scores for whatever reporters were available in that setting. As with any proportion score, it only matters *how many* symptoms were endorsed and not *which* symptoms were endorsed. Given that the binary items were coded 0 (absent) and 1 (present), the proportion score represents the average number of symptoms positively endorsed (e.g., a score of .33 reflects that 3 of the 9 symptoms were endorsed).

Importantly, the proportion scores entirely omit item characteristics (e.g., severity of symptoms), external covariates (e.g., age, treatment condition, study membership), or differential reporter perspective (child vs. parent vs. teacher). To include these potentially informative influences on score calculation, we also computed factor scores using the fully conditional TFM described earlier.[2] All models were estimated using Mplus Version 8 (Muthén & Muthén, 2017) with a logit link function and maximum likelihood with numerical integration. The estimated scoring model corresponds to that shown in Figure 1 but with the omission of the distal outcome variable; we use the distal outcome in subsequent modeling. The common factor was scaled to have a conditional latent mean and variance of 0 and 1, respectively. The child-perspective factor mean and variance were also scaled at 0 and 1, and the mean and variance for the parent- and teacher-perspective factors were freely estimated. All factor loadings were freely estimated but were equated over items (e.g., items 1, 10, and 19 were estimated and equated on the common factor, and separately estimated and equated on the perspective factors, and so on). Similar patterns of equality constraints were placed on the item intercepts as well.[3] Finally, the common factor was regressed on the set of exogenous covariates. Upon final convergence, factor scores were estimated for the common factor using standard likelihood-based procedures to obtain *expected a posteriori* (EAP) estimates.

Mediation models. To examine scores as they might be used in practice, we then estimated a manifest variable path model shown in Figure 2. Note that the exogenous variables and distal outcomes are the same as in Figure 1, but we have replaced the entire latent factor structure with a single estimated manifest score as described

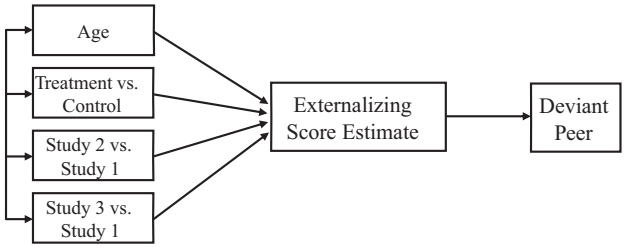

Figure 2. Path Diagram of Simulated Mediational Model Using Manifest Score Estimates.
Note. "Externalizing Score Estimate" represents the seven different estimated manifest scores presented in Table 2 with a separate model fitted to each.

above (proportion score or TFM factor score). Of key interest to us here is the mediated effect of treatment condition on deviant peer associations via child externalizing behavior estimated using the two different scoring methods across the three experimental settings.

Simulation Study: Results

The data generation, model estimation, and score calculation produced a massive amount of empirical data and subsequent results, and here we focus on a small subset of findings most relevant to our motivating questions. Specifically, we examine the relation among various score estimates across the three experimental settings, and we evaluate the performance of these score estimates in the subsequent mediational model.

Score Estimates

There were a total of eight score estimates considered: the true (population) factor score, three TFM-based score estimates (one estimated for each setting), three proportion score estimates (also one estimated for each setting), and one proportion score based on a single perspective drawn from a single study (to reflect a traditional single-reporter design). Sample statistics and correlations among score estimates are presented in Table 1.

Bivariate correlations are consistent with predictions based on psychometric theory and prior simulation results (Curran et al., 2016; Curran et al., 2018). That is, the true factor score correlates most highly with the TFM score when all reporters are present in Setting 1 ($r = .81$) and this drops modestly with partially missing reporters in Settings 2 and 3 ($r = .76$ and $r = .77$, respectively). Also as expected, the correlations between the true score and the proportion scores were markedly lower. The correlation between true scores and proportion scores when all reporters are represented in all studies (Setting 1) is $r = .78$; this decreases to $r = .65$ and $r = .69$ when reporters are partially missing (in Settings 2 and 3, respectively). Finally, the correlation between the true factor score and the proportion score for a single perspective in a single study (reflecting a typical single sample, single reporter design) is $r = .65$.

It is also interesting to consider the relations among the score estimates themselves. The correlations among the three TFM estimates range from .82 to .93, among the four proportion scores range from .47 to .81, and among the TFM estimates and the proportion scores range from .65 to .95. The lowest estimated score correlation is between the proportion score for Setting 2 and the proportion score for Setting 3 ($r = .47$), and the highest is between the TFM score and proportion score for Setting 1 ($r = .95$).

When summarizing over all correlations, two conclusions are evident. First, it is clear that the TFM score estimates reflect greater correspondence to the true scores than do the proportion scores across all three settings, particularly when there is partially missing data stemming from the simulated use of IDA. Second, the proportion scores show modest to large correlations with the corresponding TFM estimates, with some correlations being quite high. Indeed, the TFM scores and proportion scores for Setting 1 (all three reporters available in all three studies) correlate .95, a value that suggests these two score estimates might very well be interchangeable. This prompts the logical question as to whether the TFM is worthwhile when the resulting factor score correlates so highly with a score that simply represents an unweighted mean of endorsed items. However, to make a fully informed decision, we must heed Tucker's (1971) admonition to study scores as they are used in practice; we thus must move beyond simple inter-score correlations and examine how these estimates perform in subsequent modeling.

Mediational Model Results

We next considered each score as a mediator in a simple path analysis to test the extent to which hypothetical externalizing symptomatology (reflected in the score estimate) mediates the effect of the prevention condition in the prediction of subsequent delinquent peer affiliations. To test this, we estimated manifest variable path models that consisted of four exogenous covariates (representing treatment, age, and two dummy variables comparing Study 2 to Study 1 and Study 3 to Study 1), one mediator (the score estimate representing externalizing symptomatology), and one distal outcome (a continuous measure reflecting delinquent peer affiliation). We calculated the standardized mediated effect of treatment on delinquent peer via externalizing using standard methods of decomposition of effects (MacKinnon, 2012). We focused on the standardized estimate because the TFM and proportion scores are scaled to different metrics and standardization normalizes these to a shared reference. Finally, we computed the relative bias associated with each score estimate (computed as 100 times the observed estimate minus the true estimate divided by the true estimate) as a measure of effect size.

To begin, using the true generated factor score, there was a negative relation between prevention condition and externalizing symptomatology (standardized $\beta = -.33$), indicating that exposure to the prevention reduced externalizing symptomatology; further, there was a positive relation between externalizing and peer affiliation (standardized $\beta = .45$), indicating that higher levels of externalizing symptomatology were associated with greater associations with deviant peers. As such, there was a negative mediated effect of prevention condition on peer affiliation via externalizing ($-.149$), reflecting that externalizing mediated the effect of treatment on peer affiliation. We take this standardized value of $-.149$ as our population mediated effect as it is obtained from the population-generating model using true factor scores.

Next, standardized mediated effects were obtained using each available score estimate and these are presented in Table 2 (including point estimates, standard errors, and relative bias). Results show that the TFM score estimates recovered the true mediated effects exceedingly well. All three standardized estimates were nearly equal (ranging from $-.144$ to $-.147$) with relative biases approaching zero; indeed, the largest relative bias was -3.4% for the TFM

Table 1. Correlations Among True Scores, Trifactor Scores, and Proportion Scores Across Three Simulated Experimental Settings.

Variable	True score	Trifactor setting 1	Trifactor setting 2	Trifactor setting 3	Proportion setting 1	Proportion setting 2	Proportion setting 3	Proportion single study
True score	1.0							
Trifactor score setting 1	0.81	1.0						
Trifactor score setting 2	0.76	0.93	1.0					
Trifactor score setting 3	0.77	0.94	0.82	1.0				
Proportion score setting 1	0.78	0.95	0.90	0.89	1.0			
Proportion score setting 2	0.65	0.79	0.85	0.63	0.81	1.0		
Proportion score setting 3	0.69	0.84	0.68	0.89	0.87	0.47	1.0	
Proportion score single study	0.65	0.80	0.86	0.73	0.78	0.84	0.68	1.0
Mean	0	.013	−.003	−.004	.564	.456	.563	.348
SD	1	.829	.774	.785	.192	.230	.224	.242
range (min, max)	(−4.7, 4.2)	(−2.7, 2.5)	(−2.4, 2.4)	(−2.3, 2.5)	(0, 1)	(0, 1)	(0, 1)	(0, 1)

Note. Estimates are based on $n = 5{,}000$ simulated cases generated within each of three hypothetical studies resulting in a total $n = 15{,}000$.

Table 2. Standardized Mediated Effects of Treatment on Distal Outcome via the Estimated Score Across Three Simulated Experimental Settings.

Score type	Standardized mediated effect	Standard error	Relative bias
True factor scores	−.1493	.0039	—
Trifactor score estimates setting 1	−.1454	.0037	−2.61
Trifactor score estimates setting 2	−.1443	.0038	−3.35
Trifactor score estimates setting 3	−.1471	.0038	−1.47
Proportion score estimates setting 1	−.0906	.0032	−39.32
Proportion score estimates setting 2	−.0634	.0027	−57.54
Proportion score estimates setting 3	−.0690	.0027	−53.78
Proportion score estimates child report study 1	−.0537	.0045	−64.03

Note. Estimates are based on $n = 5{,}000$ simulated cases generated within each of three hypothetical studies resulting in a total $n = 15{,}000$.

score in Setting 2. The TFM estimates thus resulted in a standardized mediated effect that was only trivially smaller than the true population effect across all three settings and we consider these results to reflect no meaningful bias.

However, although the correlations among the TFM and the proportion scores were quite high (ranging into the .90's), the mediated effects associated with the proportion scores are substantially biased across all settings. Standardized point estimates for the proportion scores across the three settings ranged from −.06 to −.09 with associated relative bias values ranging from −39% to −57%. This reflects that, when using the proportion scores to capture the latent construct of externalizing symptomatology, the estimated mediated effect of the treatment on the outcome was underestimated by approximately one-half of the true population value. Worse still, the proportion score that reflected a typical single-sample, single-reporter design resulted in a point estimate of −.054 that represented an underestimation of the true mediated effect by more than 60%. This is a striking amount of bias for a score that is based on precisely the same observed data as that used in the TFM score estimates but is computed as an unweighted mean of items instead of a covariate-informed latent factor estimate.

Summary

To summarize thus far, TFM score estimates strongly correlated with their true score counterparts, whereas the correlation between proportion scores and true scores was notably lower. As expected, the presence of additional reporters in one or more contributing studies in the IDA led to better true score recovery. Interestingly, although the TFM scores were moderately to strongly correlated with the proportion scores, with one correlation equal to .95 (for Setting 1), the TFM and proportion scores led to substantially different recovery of the true mediated effects. Whereas the mediated effect was recovered with virtually no bias when using the TFM scores, this same effect was underestimated by 50% to 60% using the proportion scores obtained from precisely the same data.

These results are important given we are seeking to estimate a known population value, but we must also consider the performance of score estimates using real multiple-reporter data.

Real Data: Method

We next fit TFMs to real data drawn from a research project designed to study friendship dynamics in young adults enrolled in college. This represents a true dual-reporter situation in a single-study design in which we obtain target and friend reports of the target's own behavior. Although the real data example does not include a prevention component, it is similar in design to how such a study might be implemented within a college setting and demonstrate the broader generality of the TFM.

Participants

Data were collected as part of the Millennial Friendship Study conducted at University of North Carolina at Chapel Hill. The subsample of interest was comprised of 359 pairs of college student friend dyads (for a total 718 individual participants). Inclusion criteria for target participants were being between 18 and 26 years of age who were currently enrolled as a student at UNC and who reported any alcohol use in the past year at the point of initial screening. Peer participants had to be over age 18 and not also a target in the study. Of the 359 dyads, 67% of the targets were female, 66% of the peers were female, and 79% were same-sex.

Procedures

Participants were sent an email containing a description of the study and a link to a Qualtrics-based prescreen survey. The prescreen

evaluated the target's eligibility and, for qualifying individuals, allowed them to nominate up to four friends to participate in the study as their peer. The first-choice friend received a prescreen survey of their own to evaluate eligibility. If the peer was eligible, both the target and the peer were invited to enroll in the study. All participants completed consent procedures and a computerized battery of surveys and received a US$20 incentive.

Measures

Depressive symptoms were assessed using both target and peer report on the 13-item Short Mood and Feelings Questionnaire (SMFQ; Angold et al., 1995). The SMFQ has high internal consistency ($\alpha = .85$) and correlates moderately with the Diagnostic Interview Schedule for Children and the adult Clinical Interview Schedule–Revised Form (Angold et al., 1995; Turner et al., 2014). Target participants responded to 13 statements that reflected symptoms of depression having occurred in the past 12 months. The scale was modified so that the peers responded with respect to the target rather than themselves. Participants responded using the original SMFQ response scale (0 = *not true*, 1 = *sometimes*, 2 = *true)*; for the demonstration, responses were dichotomized based on presence or absence of the symptom (0 = *not true*, 1 = *sometimes true or true*).

Past year depression for the target was also assessed using a binary screener item drawn from the Structured Clinical Interview for DSM-V: "In the past year, has there been a period of time when you were feeling depressed or down most of the day, nearly every day?" (0 = *no*, 1 = *yes*).

To evaluate the quality of the relationship between the target and the peer, the target completed the 30-item Network of Relationships Inventory: Social Provision Version (Buhrmester & Furman, 2008). We used the 3-item *intimate disclosure* subscale that measures how frequently individuals share private, sensitive, or personal information with their friends. All 3 items were rated on a 5-point Likert-type scale ranging from 0 (*little to none*) to 4 (*the most possible*), and the average of the 3 items was used in the analysis.

Finally, both targets and peers self-reported their own biological sex (0 = *female*, 1 = *male*).

Real Data: Results

We fit a series of TFMs to the MFS data using Mplus Version 8 (Muthén & Muthén, 2017). Full-information maximum likelihood was used to calculate likelihood ratio tests and subsequent score estimates. We followed the model building procedures described by Bauer et al. (2013).

Unconditional TFM

We began with the estimation of an unconditional TFM fit to the MFS data described above; this model is of the same structure as that presented in Figure 1, but here the exogenous variables are NRI, sex, and 12-month depression, there is one common factor for depression, two reporter factors for target and peer, and 13 specific factors, one for each item. To identify the model, the factor loadings and intercepts for the first item on each of the perspective factors (i.e., item 1 for the target, item 1 for the peer) were set to equality. Further, the loadings for each item pair that defined the specific factors were also set to one and the variances were freely estimated. The common factor and the target perspective factor were constrained to have a mean of zero and variance of one, while the mean and variance of the peer perspective factor were freely estimated.

Next, we imposed equality constraints on the item intercepts and factor loadings to determine whether factorial invariance held across peers and targets. Briefly, establishing factorial invariance ensures that informants are responding to the items in the same way and allows for parameter estimates to be meaningfully compared. Equality constraints were first added for all factor loadings across reporter to test for metric invariance. The likelihood ratio test comparing this model to the unrestricted model was nonsignificant, $\Delta\chi^2(25) = 20.66$, $p = .71$, thus supporting the more restricted metric invariance model. Intercepts were then constrained to be equal across perspectives, which resulted in a significant likelihood ratio test, $\Delta\chi^2(12) = 52.11$, $p < .001$, therefore not supporting the more constrained scalar invariance model. Using parameter estimates from the metric invariance model, a series of partial invariance models in which single intercepts were freed were fit. A model in which seven of the intercepts were invariant and six were free to vary resulted in a nonsignificant likelihood ratio test ($\Delta\chi^2(6) = 11.54$, $p = .07$). For five of the six intercepts that were free to vary, the intercept was lower for the target than it was for the peer, meaning that the item was easier for the peer to endorse. For one of the items, the threshold was higher for the target, meaning that it was easier for the target to endorse. These constraints are sufficient to expand this model to include the set of covariates.

Conditional TFM

We included the impact of three target-specific predictors on a subset of the latent factors: target report of biological sex, intimate disclosure, and past-year depression. The common factor was regressed on all three measures and target's perspective factor was regressed on just past-year depression; this latter effect tested the extent to which the target's perspective on the 13 depression items itself was influenced by the target's self-reported past-year depression.

The factor loadings for both the common and perspective factors from the final conditional model were all positive and significant ($p < .05$) for all items (see Table 3 for the standardized estimates). However, the magnitude of the perspective factor loadings was larger relative to those of the items on the common factor. This suggests that the common factor contributes somewhat less to the item responses than the variation associated with the two perspectives. While the estimated mean for the peer perspective factor was lower than that of the target perspective factor ($\hat{\mu}_{\text{peer}} = -.15$), this was not statistically significant, which means that the peer did not rate the target as having significantly less depressive symptomatology than did the target themselves.

Regression estimates for the common factor showed that male targets displayed lower levels of depressive symptomatology than females ($\beta = -.30$, $p < .001$). Additionally, targets who endorsed the past-year measure of depression displayed higher levels of depressive symptomatology ($\beta = .32$, $p < .001$). In contrast, the measure of intimate disclosure did not significantly predict the common depression factor. Finally, the regression of the target perspective factor on past-year depression was positive and significant ($\beta = .43$, $p < .001$), suggesting that individuals who reported past-year depression provided systematically higher ratings of their

Table 3. Standardized Intercept and Factor Loading Estimates From the Final Conditional Trifactor Model for the $N = 359$ Dyads From the Millennial Friendship Study.

			Factor loading		
Item	Intercept		Common	Perspective	Specific
1. Miserable or unhappy	−0.40		0.35	0.74	0.00
2. Didn't enjoy anything	0.61		0.34	0.79	0.23
3. So tired sat around and did nothing	−0.27		0.16	0.61	0.26
4. Was very restless	−0.26	−0.01	0.20	0.61	0.32
5. Felt no good any more	0.46	0.68	0.58	0.65	0.03
6. Cried a lot	0.32	0.55	0.50	0.40	0.60
7. Hard to think properly or concentrate	−0.02		0.27	0.64	0.21
8. Hated self	0.80		0.49	0.67	0.14
9. Was a bad person	0.81		0.50	0.58	0.15
10. Felt lonely	−0.22	0.01	0.37	0.64	0.21
11. Thought nobody loved them	0.86		0.60	0.54	0.13
12. Never be as good as other people	0.18	0.53	0.60	0.48	0.07
13. Did everything wrong	0.71	0.62	0.64	0.54	0.04

Note. invariant intercepts are indicated by a single shared value; non-invariant intercepts are indicated by the first value for the target report and the second value for the peer report

own depressive symptoms relative to those who did not report past-year depression.

Scoring

Finally, we estimated factor scores (as maximum likelihood EAP scores) for the common factor assessing the target's depression based on the conditional model described above. These scores reflect individual variability in target depression while incorporating the effects of reporter perspective and specific item characteristics. Unlike simple sum or proportion scores, EAPs up-weight items that are more strongly related to the factor and down-weight those that are less strongly related and thus do not simply consider how many items were endorsed, but *which* items were endorsed. To highlight this, we computed simple sum scores in two ways that are commonly used in practice: one based just on the sum of the target's 13-item responses (thus ignoring the second reporter) and one based on the combined mean of the target's and peer's 13-item responses. The correlations between the EAPs and the target-only and the combined mean scores were .69 and .83, respectively, while the two sum scores correlated .84 with one another. When compared to the TFM scores, there is a loss of information when combining the observed items by taking simple sums.

But the advantages of the TFM scores relative to the sum scores extend one step further. Figure 3 presents the distribution of common factor scores assessing target depression at each possible sum score based on the target's own ratings. This plot clearly shows that for any given sum score there exists an *entire distribution* of factor score values reflecting substantially enhanced individual variability in the assessment of target depression. For example, 33 targets reported a sum score equal to 5.0 (thus all having precisely the same sum score) yet there is an entire distribution of factor scores for these same respondents. Further, for the 33 targets receiving a self-report score of 5.0, only 3 of the 33 peer reporters *also* reported a 5.0; indeed, peer sum scores ranged from 0 to 11 with a median of 3.0 and 21 of the 33 (63%) reported sum scores *lower* than 5.0. The TFM clearly adds substantial variability into the scores by combining both target and peer report that could then be capitalized upon in subsequent modeling. These TFM scores could then be used in subsequent model fitting and hypothesis testing, but we do not pursue this further here.

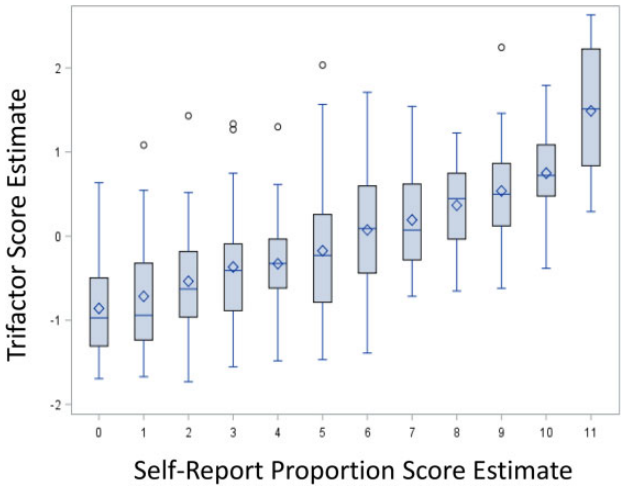

Figure 3. Bivariate Box Plots of Self-Report Proportion Scores and Trifactor Scores for $N = 359$ Dyads From the Millennial Friendship Study.

Discussion

Conducting rigorous empirical research in the prevention sciences is a breathtakingly challenging endeavor. Empirically evaluating a prevention or intervention program touches on a host of critically important methodological and quantitative elements including design, sampling, assessment, and valid causal inference in the prediction of change over time (e.g., Brown, 1993; 2003). A fundamental issue that underlies all empirical studies in the prevention sciences is that of measurement and scoring. Obtaining accurate numerical measures of our underlying theoretical constructs is critical, given that this process provides the central unit-of-analysis for making inferences about the etiological mechanisms under study. Over a century of research has resulted in advanced and comprehensive psychometric approaches for obtaining observed measures of often unobservable phenomena such as depression, anxiety, temperament, and self-esteem. However, these methods are not without limitations.

Chief among these is the well-established fact that making empirically based inferences based solely on a single reporter leads to an inextricable confound between the target being assessed and the perspective of the individual reporter. This is particularly salient if the reporter and the target are one-in-the same; that is, *self-report*. Although a deep and thoughtful appreciation of both the theoretical and methodological implications of single versus multiple reporter has been evidenced for many decades (see, e.g., De Los Reyes, 2011; De Los Reyes et al., 2013), only recently have comprehensive analytic methods become available for evaluating multiple reporter data in a psychometrically principled way. One such novel

method is the TFM (Bauer et al., 2013). The TFM is a logical extension of the traditional bifactor model (Gibbons et al., 2007; Reise et al., 2011) that incorporates assessments obtained from two or more independent reporters. Here we used simulated and real data to study the TFM for analyzing multiple reporter data under conditions commonly encountered in IDA within the prevention sciences and related disciplines. Although our focus is on prevention research applications, these methods generalize to a much broader set of empirical applications in the behavioral, health, and educational sciences.

The results of our simulation study were largely consistent with theoretical expectations, although the magnitude of bias present when using the traditional scale scores surprised even us. We designed our simulation to reflect an IDA in which data from three separate prevention studies were pooled and examined, each of which had different combinations of reporters. We considered three settings: one where all reporters were present for all studies, one where all reporters were present for one study but only a single reporter was present for the other two studies, and one where each study had a unique pairing of reporter. Using these simulated data, we then computed score estimates both from the TFM and using traditional proportion scores of available items for whatever reporters were present in each condition. The TFM-based scores clearly demonstrated higher correlations with the true underlying score compared to the proportion scores. However, the score estimates themselves correlated highly with one another; indeed, the correlation between the TFM and proportion scores for the complete-case setting was .95. This makes one wonder if the added complexity associated with the TFM is worthwhile. However, this wonderment is quickly put to rest when considering the performance of scores when taken to subsequent analyses.

More specifically, we compared the proportion score and TFM factor scores (taken to represent externalizing symptomatology) as mediators in the relation between a hypothetical treatment condition and a distal outcome (taken to represent affiliations with deviant peers) across three experimental settings. Whereas the TFM-based scores recovered the population-mediated effect almost perfectly, the traditional proportion scores led to drastic underestimates of the mediated effect, at times exceeding 50%. Further, the worst recovery of all was observed with proportion scores based on a single reporter drawn from a single study, a condition that is very common in applied research. Two conclusions are clear. First, using multiple reporter data when available improves the recovery of the mediated effect, at least under the conditions studied here. Second, if an incorrect measurement model is used to obtain scale scores (i.e., proportion scores), and these scale scores are then used as a mediator in a subsequent model, then the mediated treatment effect can be significantly underestimated relative to its population value. This is a striking reduction in effect size and is one that can be mitigated by applying a different psychometric model to the very same data.

To expand our study of artificial data, we applied the TFM to a large sample of friendship dyads that were collected as part of a study of college student friendships. Both the target and their peer each reported on 13 items assessing depressive symptomatology of the target. Although this study design did not have an IDA component, it is similar in form to what might be used in other prevention-based or prevention-like studies. Exogenous covariates were included in the TFM as predictors of both the general depression factor and the target's own perspective factor. Results showed that although all items loaded positively on both the common and the two perspective factors, the items were more strongly associated with the individual perspectives. This suggests that a major component of depressive symptomatology is in the eye of the beholder; that is, how each reporter perceives the depression items is stronger than the pooled underlying component believed to have given rise to the items (see De Los Reyes et al., 2013, for a detailed discussion of these issues). Further, factor score estimates were obtained for the target's depression that combined the target and peer reports while controlling for the reporter-specific perspectives. The TFM provided psychometrically superior scores to those obtained using traditional proportions of endorsed items and these are in turn available for subsequent analysis.

Potential Limitations

Despite the strengths of our findings, there of course remain certain limitations. First, we used a simulation design in which three large samples of data were generated and examined. This has the distinct advantage of studying more stable aspects of model estimation and score calculation, but it does not allow for a systematic examination of sampling variability. Our prior simulations strongly suggest that this pattern of results would hold across smaller and more realistic sample sizes (Curran et al., 2016; Curran et al., 2018), but future studies could expand our design to allow for a closer examination of the influence of variability introduced by small sample size. Second, the simulation could be further extended to include a variety of factors that are common in prevention and intervention studies including additional covariates, different item response options, and different item sets associated with each reporter. Further, although we examined the direct effects of the exogenous covariates on the latent factors, we did not consider more complex forms of DIF (Holland & Wainer, 2012). Our own prior research on scoring single factor models suggests that DIF effects play an important role in modeling and scoring (Curran et al., 2016; Curran et al., 2018) and these would be expected to hold in the TFM as well. Finally, our work focused on a design corresponding to a single point in time. This of course would never occur in an actual prevention study, and a key issue in need of further attention is how the TFM can best be extended to repeated measures assessments.

Summary

Both theory and empirical results establish that for many research applications there are significant potential advantages to including assessments drawn from two or more reporters. Historically, the disagreement between two reporters has often been considered "bias" that contaminate the estimation of associated scores. However, these differences can alternatively be thought of as "perspective" and through the TFM can be made a formal part of the psychometric model. Our simulation results and analysis of real data lead to two clear conclusions. First, under the conditions studied here, the inclusion of multiple reporter assessments when available is superior to using the assessment obtained from a single reporter alone; this conclusion holds whether used in a single study design or when combining multiple studies within an IDA framework. Second, across all conditions studied here, the TFM-based factor scores were psychometrically superior to those obtained using traditional proportion score calculations. This same finding has been unambiguously evident in our prior work focused on

scoring with single factor models (Curran et al., 2014; Curran et al., 2016; Curran et al., 2018). We thus recommend that the recent advances in scoring based on covariate-informed factor estimation be closely considered for use in practice.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by R01DA034636 (D. J. Bauer, Principal Investigator) and the National Institute on Drug Abuse.

ORCID iD

Patrick J. Curran https://orcid.org/0000-0002-5772-5120
A. R. Georgeson https://orcid.org/0000-0002-6426-9258

Notes

1. Although our simulations are intentionally based on large sample sizes with the intent of minimizing the effects of sampling variability, these results would be fully expected to reflect smaller and more realistic sample sizes more commonly encountered in practice.
2. One additional scoring model that could be considered is the trifactor model that does not include covariates. However, our prior simulation findings with single reporter data clearly indicate that such a model is superior to proportion scores and inferior to covariate-informed scores, so we do not replicate these findings here (see Curran et al., 2016; Curran et al., 2018).
3. See Bauer et al. (2013) for details about the imposing and testing of equality constraints in the trifactor model.

References

Achenbach, T. M. (1999). The child behavior checklist and related instruments. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed., pp. 429–466). Lawrence Erlbaum Associates Publishers.

Achenbach, T. M. (2011). Commentary: Definitely more than measurement error: But how should we understand and deal with informant discrepancies? *Journal of Clinical Child & Adolescent Psychology*, *40*, 80–86. https://doi:10.1080/15374416.2011.533416

Achenbach, T. M., & Edelbrock, C. S. (1981). Behavioral problems and competencies reported by parents of normal and disturbed children aged four through sixteen. *Monographs of the Society for Research in Child Development*, *46*, 1–82.

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, *131*, 361–382. https://doi:10.1037/0033-2909.131.3.361

Angold, A., Costello, E. J., Messer, S. C., & Pickles, A. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research*, *5*, 237–249.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Bandalos, D. L., & Finney, S. J. (2018). Factor analysis: Exploratory and confirmatory. In G. R. Hancock, R. O. Mueller, & L. M. Stapleton (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 110–134). Routledge.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. https://doi:10.1037/met0000077.supp (Supplemental)

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, *18*, 475–493. https://doi:10.1037/a0032475

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*, 101–125. https://doi:10.1037/a0015583.supp (Supplemental)

Boyle, M. H., & Pickles, A. R. (1997). Influence of maternal depressive symptoms on ratings of childhood behavior. *Journal of abnormal child psychology*, *25*, 399–412.

Brown, C. H. (1993). Statistical methods for preventive trials in mental health. *Statistics in Medicine*, *12*, 289–300.

Brown, C. H. (2003). Design principles and their application in preventive field trials. In W. J. Bukoski & Z. Sloboda (Eds.), *Handbook of drug abuse prevention: Theory, science, and practice* (pp. 523–540). Kluwer Academic/Plenum Press.

Buhrmester, D., & Furman, W. (2008). *The network of relationships inventory: Relationship qualities version*. Unpublished measure, University of Texas at Dallas.

Collins, L. M., & Flaherty, B. P. (2006). Methodological considerations in prevention research. In S. Zili & W. J. Bukoski (Eds.), *Handbook of drug abuse prevention* (pp. 557–573). Springer.

Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods*, *14*, 77–80.

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 827–844. https://doi:10.1080/10705511.2016.1220839

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling*, *25*, 860–875. https://doi:10.1080/10705511.2018.1473773

Curran, P. J., Howard, A. L., Bainter, S. A., Lane, S. T., & McGinley, J. S. (2014). The separation of between-person and within-person components of individual change over time: A latent curve model with structured residuals. *Journal of Consulting and Clinical Psychology*, *82*, 879–894. https://doi:10.1037/a0035297

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*, 81–100. https://doi:10.1037/a0015914

De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *40*, 1–9. https://doi:10.1080/15374416.2011.533405

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*, 123–149.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*, 1–11.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Gibbons, R. D., Bock, R. D., Hedeker, D., Wiess, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19. https://doi:10.1177/0146621606289485

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

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

MacKinnon, D. (2012). *Introduction to statistical mediation analysis*. Routledge.

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Clinical applications of integrative data analysis. *Annual Review of Clinical Psychology*, *9*, 61–89.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Author.

Reise, S. P., Ventura, J., Keefe, R. S. E., Baade, L. E., Gold, J. M., Green, M. F., Kern, R. S., Mesholam-Gately, R., Nuechterlein, K. H., Seidman, L. J., & Bilder, R. (2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological Assessment*, *23*, 245–261. https://doi:10.1037/a0021501

Renk, K. (2005). Cross-informant ratings of the behavior of children and adolescents: The "gold standard". *Journal of Child and Family Studies*, *14*, 457–468. https://doi:10.1007/s10826-005-7182-2

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Erlbaum.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurement of educational products. In S. A. Courtis (Ed.), *The measurement of educational products* (17th Yearbook of the National Society for the Study of Education, *Pt. 2*. pp. 16–24). Public School.

Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika*, *36*, 427–436. https://doi:10.1007/BF02291367

Turner, N., Joinson, C., Peters, T. J., Wiles, N., & Lewis, G. (2014). Validity of the short mood and feelings questionnaire in late adolescence. *Psychological Assessment*, *26*, 752–762.

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