Published in final edited form as:
Psychol Methods. 2013 December ; 18(4): 475–493. doi:10.1037/a0032475.

A Tri-Factor Model for Integrating Ratings Across Multiple Informants

Daniel J. Bauer,
The University of North Carolina at Chapel Hill

Andrea L. Howard,
The University of North Carolina at Chapel Hill

Ruth E. Baldasaro,
The University of North Carolina at Chapel Hill

Patrick J. Curran,
The University of North Carolina at Chapel Hill

Andrea M. Hussong,
The University of North Carolina at Chapel Hill

Laurie Chassin, and
Arizona State University

Robert A. Zucker
University of Michigan

Abstract

Psychologists often obtain ratings for target individuals from multiple informants such as parents or peers. In this paper we propose a tri-factor model for multiple informant data that separates target-level variability from informant-level variability and item-level variability. By leveraging item-level data, the tri-factor model allows for examination of a single trait rated on a single target. In contrast to many psychometric models developed for multitrait-multimethod data, the tri-factor model is predominantly a measurement model. It is used to evaluate item quality in scale development, test hypotheses about sources of target variability (e.g., sources of trait differences) versus informant variability (e.g., sources of rater bias), and generate integrative scores that are purged of the subjective biases of single informants.

When measuring social, emotional, and behavioral characteristics, collecting ratings from multiple informants (e.g., self, parent, teacher, peer) is widely regarded as methodological best practice (Achenbach, McConaughy & Howell, 1987; Achenbach et al., 2005; Renk, 2005). Similarly, the use of multisource performance ratings (e.g., by supervisors, peers, and subordinates, a.k.a., 360-degree assessment) is considered an optimal method for evaluating

Correspondence concerning this research should be addressed to: Daniel Bauer, Department of Psychology (CB#3270), University of North Carolina, Chapel Hill NC 27599-3270; dbauer@email.unc.edu.

The content is solely the responsibility of the authors and does not represent the official views of the National Institute on Drug Abuse, National Institute on Alcohol Abuse and Alcoholism, or the National Institutes of Health.

job performance (Conway & Huffcutt, 1997; Lance et al., 2008). Because each informant observes the target within a specific relationship and context, it is desirable to obtain ratings from multiple informants privy to different settings to obtain a comprehensive assessment, particularly when behavior may vary over contexts. Further, any one informant's ratings may be compromised by subjective bias. For instance, depressed mothers tend to rate their children higher on psychopathology than do unimpaired observers, in part because of their own impairment and in part because of an overly negative perspective on their children's behavior (Boyle & Pickles, 1997; Fergusson, Lunskey & Horwood, 1993; Najman, 2000; Youngstrom, Izard & Ackerman, 1999). In principle, researchers can better triangulate on the true level of the trait or behavior of interest from the ratings of several informants. The idea is to abstract the common element across informants' ratings while isolating the unique perspectives and potential biases of the individual reporters.

Although there is widespread agreement that collecting ratings from multiple informants has important advantages, less agreement exists on how best to use these ratings once collected. The drawbacks of relatively simple analysis approaches, such as selecting an optimal informant, conducting separate analyses by informant, or averaging informant ratings, have been discussed at some length (e.g., Horton & Fitmaurice, 2004; Kraemer, Measelle, Ablow, Essex, Boyce & Kupfer, 2003; van Dulmen & Egeland, 2011). In contrast to these approaches, psychometric models offer a more principled alternative wherein the sources of variance that contribute to informant ratings can be specified and quantified explicitly.

Several psychometric modeling approaches have already been proposed for analyzing multiple informant data. Reviewing these approaches, we identify the need for a new approach that is more expressly focused on issues of measurement. We then propose a new measurement model for multiple informant data, which we refer to as the tri-factor model (an adaptation and extension of the bi-factor model of Holzinger and Swineford, 1937). The tri-factor model stipulates that informant ratings reflect three separate sources of variability, namely the common (i.e., consensus) view of the target, the unique (i.e., independent) perspectives of each informant, and specific variance associated with each particular item. Correlated (non-unique) perspectives can also be accommodated by the model when some informants overlap more than others in their views of the target, albeit at some cost to interpretability. We demonstrate the advantages of the tri-factor model via an analysis of parent reports of childrens' negative affect. Finally, we close by discussing directions for future research.

Psychometric Models for Multiple Informant Data

A number of different psychometric models for multiple informant data have appeared in the literature. As noted by Achenbach (2011), most of these models represent adaptations of models originally developed for the analysis of multitrait-multimethod (MTMM) data (Campbell & Fiske, 1959). Such data are often collected with the goal of assessing convergent and discriminant validity, as evidenced by the magnitude of inter-trait correlations after accounting for potential method-halo effects which could otherwise distort these correlations (Marsh & Hocevar, 1988). Quite often, the "methods" are informants, and the goal is then to isolate trait variability from rater effects and measurement error (e.g., Alessandri, Vecchione, Tisak, & Barbaranelli, 2011; Barbaranelli, Fida, Paciello, Di Giunta, & Caprara, 2008; Biesanz & West, 2004; Eid, 2000; Pohl & Steyer, 2010; Woehr, Sheehan, & Bennett, 2005).[1]

One of the earliest proposed psychometric models for MTMM data is the Correlated Trait/ Correlated Method model (CTCM; Marsh & Grayson, 1995; Kenny, 1979; Widaman, 1985). Given multiple trait ratings from each of multiple informants, the CTCM model is structured

so that each rating loads on one trait factor and one method (i.e., informant) factor. Across methods, all ratings of the same trait load on the same trait factor and, across traits, all ratings made by the same method load on the same method factor. Trait factors are permitted to correlate, as are method factors, but trait and method factors are specified to be independent of one another.

Although the CTCM structure nicely embodies many of the ideas of Campbell and Fiske (1959), it has proven difficult to use in practice. In particular, the CTCM model often fails to converge (Marsh & Bailey, 1991), and suffers from potential empirical under-identification (Kenny & Kashy, 1992). As a more stable alternative, Marsh (1989) advocated use of the Correlated Trait/Correlated Uniqueness model (CTCU; see also Kenny, 1979). The CTCU model maintains correlated trait factors; however, method factors are removed from the model. To account for the dependence of ratings made by the same informant, the uniquenesses (i.e., residuals) of the measured variables are intercorrelated within informant.

Due to the lack of explicit method factors, the CTCU model has been criticized for failing to separate method variance from error variance, potentially leading to underestimation of the reliability of the measured variables (Eid, 2000; Lance et al., 2002; Pohl & Steyer, 2010). Additionally, whereas the CTCM model permits correlated method factors, in the CTCU model uniquenesses are not permitted to correlate across methods. This feature of the CTCU model has been considered both a strength and a limitation. Marsh (1989) argued that if a general trait factor (i.e, higher order factor) influenced all trait ratings then correlated method factors might actually contain common trait variance. By not including across-method correlations, the CTCU model removes this potential ambiguity from the model. Yet this also implies that if methods are truly correlated (e.g., informants are not independent) then the trait variance may be inflated by excluding these correlations from the model (Conway et al., 2004; Kenny & Kashy, 1992; Lance et al., 2002).

Several additional criticisms are shared by the CTCM and CTCU models. Marsh and Hocevar (1988) and Marsh (1993) noted that CTCM and CTCU models are typically fit to scale-level data (i.e., total scores) with unfortunate implications. First, the uniqueness of each manifest variable contains both measurement error and specific factor variance, yet because only one measure is available for each trait-rater pair there is no ability to separate these two sources of variance. Thus the factor loadings may underestimate the reliability of the manifest variables to the extent that the specific factor variance is large. Second, when specific factors are correlated, as might occur when informants rate the same items, this could artificially inflate the trait variances obtained by analyzing the scale-level data. Marsh and Hocevar (1988) and Marsh (1993) thus advocated using second-order CTCM and CTCU models in which the traditional scale-level manifest variables are replaced with multiple indicator latent factors defined from the individual scale items or item parcels.

At a more fundamental level, Eid (2000) and Pohl, Steyer and Kraus (2008) critiqued the CTCM and CTCU models for being based largely on intuition, incorporating arbitrary restrictions, and having ambiguously defined latent variables. To address these shortcomings, Eid (2000) proposed the CTC(*M*-1) model. This model is defined similarly to the CTCM model except that the method factor is omitted for a “reference method” chosen by the analyst (e.g., self-reports). The remaining method factors (e.g., parent- and peer-report factors) are then conceptualized as residuals relative to the reference method. The

[1]Less often, data is obtained from multiple informants on multiple targets (e.g., siblings) for a single trait (e.g., Rowe & Kandel, 1997; Hewitt, Silberg, Neale, Eaves & Erickson, 1992; Neale & Stevenson, 1989; Simonoff et al., 1995; van der Valk, van den Oord, Verhulst & Boosma, 2001). These models are often applied to twin data with the goal of partitioning the trait variance into genetic and environmental components without the contaminating influence of informant bias (e.g., Bullock, Deater-Deckard, & Leve, 2006).

notion is that there is a true score underlying any given trait rating and when this true score is regressed on the corresponding trait rating from the reference method the residual constitutes the method effect (with an implied mean of zero and zero correlation with the reference method trait ratings). Eid et al (2003) also extended the CTC(*M*-1) model to allow for multiple indicators for each method-trait pair, and to allow for trait-specific method effects. This latter extension is important in relaxing the assumption that method effects will be similar across heterogeneous traits.

The CTC(*M*-1) model has the virtue of having well-defined factors and a clear conceptual foundation. It has, however, been criticized on the grounds that the traits must be defined with respect to a reference method and the fit and estimates obtained from the model are neither invariant nor symmetric with respect to the choice of reference method (Pohl & Steyer, 2010; Pohl, Steyer & Kraus, 2008). Eschewing the definition of method effects as residuals, Pohl et al. (2008) argued that it is better to conceptualize method factors as having causal effects on the underlying true scores of the trait ratings. As causal predictors, method factors would neither be implied to have means of zero nor to have zero correlation with the trait factors (also construed to be causal predictors). The MEcom model of Pohl and Steyer (2010) thus permits method factor means and method-trait correlations to be estimated. The MEcom model also obviates the need to select a reference method, thus allowing traits to be defined as what is common to all raters. Finally, given multiple indicators for each trait-method pair, the MEcom model retains the advantage that trait-specific method factors can be specified.

Despite these many advantages, the MEcom model also has potential limitations. First, method effects are defined as contrasts between informants (imposed through restrictions on the method factor loadings), not as the effect of a given informant. Thus the unique effects of the informants are not separated, complicating the assessment of specific sources of rater bias. Second, some of the common variance across informant trait ratings will be accounted for by the method factors. That is, trait ratings made by disparate informants will correlate not only due to the common influence of the underlying trait but also due to the fact that they are influenced by correlated methods. Although Pohl and Steyer (2010) regard such a specification as more realistic, it runs counter to the argument made by Marsh (1989) that trait factors should exclusively account for the correlations among trait ratings across different methods. Third, due to the inclusion of trait-method correlations (unique to this model), the variances of the trait ratings cannot be additively decomposed into trait, method, and error components.

In sum, a variety of psychometric models have been developed with the goal of separating trait and method variability in informant ratings, and each of these models has specific advantages and limitations. The common goal of all of these developments has been to improve the analysis of MTMM data for evaluating convergent and discriminant validity. Yet many researchers do not collect multiple informant data with the intent of examining construct validity. Often, the primary goal is simply to improve construct measurement. That is, researchers seek to generate integrated scores of the construct of interest that optimally pool information across informants who have unique access to and perspectives on the target individuals. There is then less concern with estimating inter-trait correlations and greater concern with scale development and the estimation of scores for use in substantive hypothesis testing.

The psychometric modeling approach we develop in this paper is thus designed for the situation in which one wishes to evaluate and measure a single construct for a single target based on the ratings of multiple informants, and to extract integrated scale scores for use in subsequent analysis. This situation clearly does not parallel the usual MTMM design and

hence models initially developed for MTMM data are not readily applicable.[2] Bollen and Paxton (1998) demonstrated, however, that one can often decompose informant ratings into variation due to targets (reflected in all informants' ratings, such as trait variation) and variation due to informants (reflecting informants' unique perspectives, contexts of observation, and subjective biases) even when data do not conform to the traditional structure of an MTMM design. Bollen and Paxton noted that this endeavor is greatly facilitated by the availability of multiple observed indicator variables for the trait. In the present paper, we draw upon and extend this idea to present a novel model for evaluating common and unique sources of variation in informant ratings.

Taking advantage of recent advances in item factor analysis, our model is specified at the item level, leveraging the individual item responses as multiple indicators of the trait. We stipulate a tri-factor model for item responses that includes a common factor to represent the consensus view of the target, perspective factors to represent the unique views (and biases) of the informants, and specific factors for each item. It is conceptually advantageous to assume these factors are independent contributors to informants' ratings. After introducing the model, however, we shall describe instances in which one might find it useful or practically necessary to introduce correlations between subsets of factors. It is also notable that the tri-factor model does not require that a specific informant be designated as a reference (as required by the CTC(*M*-1) model), permitting the common factor to be interpreted as what is common to all informants. Nor does the tri-factor model represent informant effects via contrasts between informants (as required by the MEcom model), thus enabling the evaluation of putative sources of bias for specific informants' ratings.

The tri-factor model has three primary strengths. First, with the tri-factor model we can evaluate the extent to which individual items reflect the common factor, perspective factors, and item-specific factors. This information can be quite useful in determining which items are the most valid indicators of the trait versus those that are most influenced by the idiosyncratic views of the rater. In contrast, scale-level data do not provide any information on the validity of individual items. Further, many scales were developed through item-level factor analyses conducted separately by informant, yet such analyses cannot distinguish common factor versus perspective factor variability. Thus an item may appear to factor well with other items due to perspective effects alone even if it carries little or no information about the trait. The tri-factor model separates these sources of variability to provide a more refined psychometric evaluation of the items and scale.

A second strength of the tri-factor model is that we can include external predictors to understand the processes that influence informants' ratings. In particular, we can regress the common, perspective, and specific factors on predictor variables to evaluate systematic sources of variability in trait levels, observer effects, and specific item responses. For instance, we might expect the child of a depressed parent to have a higher level of negative affect, represented in the effect of parental depression on the common factor. In addition, a parent who is depressed may be a biased rater of his or her child's negative affect. The latter effect would be captured in the effect of parental depression on the perspective factor of the depressed parent.

Finally, a third strength of the tri-factor model is that it is expressly designed to be used as a measurement model. That is, a primary goal in fitting a tri-factor model is to obtain scores on the common factor. These scores provide an integrative measure of the characteristic of

[2]Indeed, many MTMM models are not identified for a single trait, at least when fit to scale-level data. The availability of multiple items for each method permits adaptation of some of these models to the measurement of a single trait, although we are not aware of any single-trait applications of these models in the literature.

interest that is purged of both known and unknown sources of rater bias. Scores for the perspective factors may also be of interest, for instance when research focuses on why informants evaluate targets differently. In this sense, the purpose in fitting a tri-factor model is quite different from the purpose of fitting MTMM models, which are more typically aimed at evaluating construct validity.

In sum, the tri-factor model contributes to a strong tradition of psychometric models for analyzing multiple informant data. It is designed for the purpose of generating integrated scores from item-level data across multiple informants and it enables researchers to evaluate processes that influence both the common and unique components of informant ratings. In what follows we further explicate the tri-factor model and we discuss model estimation and scoring. We then show how the model can be applied through a real-data example.

The Tri-Factor Model for Multiple Informant Data

We begin by describing the decomposition of variability in item responses that is the basis of the tri-factor model. We then describe how the model can be extended to incorporate predictors. Following our description of the model we discuss estimation and scoring.

Unconditional Tri-Factor Model

The unconditional tri-factor model consists of a set of observed ratings from each informant, which we shall refer to as item responses, and three types of latent variables. We assume that the items all measure a single, unidimensional trait (an assumption that, in practice, can be evaluated by goodness of fit testing). We designate a given item response as y_{irt} and its expected value as μ_{irt}, where i indicates the item ($i = 1, 2, \ldots I$), r indicates the rater type (e.g., self, teacher, mother; $r = 1, 2, \ldots R$), and t indicates the target ($t = 1, 2, \ldots N$). The item set and rater types are taken to be fixed, whereas targets are assumed to be sampled randomly from a broader population of interest. We shall initially assume parallel item sets across raters, but later describe the application of the model to differential item sets. The rater types may be structurally different (e.g., co-worker versus supervisor) or interchangeable (e.g., two randomly chosen co-workers), a distinction made by Eid et al. (2008) for MTMM models (see also Nussbeck et al., 2009, and the related concept of "distinguishability" as defined by Gonzalez & Griffin, 1999, and Kenny, Kashy & Cook, 2006). Latent variables represent sources of variability in the item responses across targets. The three types of latent variables are C_t, P_{rt}, and S_{it}, representing, respectively, a common factor, R unique perspective factors (one for each informant), and I specific factors (one for each item). The latent variables are assumed to be normally distributed and independent.

The structure of the tri-factor model is

$$g_i\ (\mu_{irt}) = \nu_{ir} + \lambda_{ir}^{(C)} C_t + \lambda_{ir}^{(P)} P_{rt} + \lambda_{ir}^{(S)} S_{it} \quad (1)$$

Focusing first on the left side of the equation, $g_i(\cdot)$ is a link function chosen to suit the scale of y_{irt}, such as the identity link for a continuous item or the logit or probit link for a binary item. The link function can vary across items, allowing for items of mixed scales types (Bartholomew & Knott, 1999; Bauer & Hussong, 2009; Skrondal & Rabe-Hesketh, 2004). The inverse of the link function, or $g_i^{-1}(\cdot)$ returns the expected value of the item response. For a continuous item response the expected value would be a conditional mean and for a binary item response it would be a conditional probability of endorsement (conditional on the levels of the factors).

Turning now to the right side of the equation we see the usual set up for a factor analytic model with an intercept ν_{ir} and factor loadings $\lambda_{ir}^{(C)}$, $\lambda_{ir}^{(P)}$, and $\lambda_{ir}^{(S)}$ for the three types of factors. Conditional on the factors, item responses are assumed to be locally independent. An example tri-factor model for 13 items rated by two informants, mothers and fathers, is shown in path diagram form in Figure 1. Note that each item loads on one of each type of factor: the common factor, a unique perspective factor for the informant, and a specific factor for the item.

The factors are conceptually defined and analytically identified by imposing constraints on the factor loadings and factor correlations. To start, all informant ratings are allowed to load on the common factor C_t. This factor thus reflects shared variability in the item responses across informants. It is considered to represent the consensus view of the target across informants. This consensus will reflect trait variability as well as other sources of shared variability (Kenny, 1991). Informants may observe the target in the same context, may relate to the target in similar ways, or may directly share information with one another. For instance, mothers and fathers both serve the role of parent and both observe their child's behavior principally within the home environment. Parents may also communicate with each other about their children's problem behavior. Given data from both parents, as in Figure 1, C_t would represent the common parental view of the child's behavior, whatever its sources (trait, context, common perspective, mutual influence). If the goal of fitting the model is to isolate general trait variability within C_t then other sources of shared variability should be minimized in the research design phase, for instance by selecting dissimilar informants who observe the child in different contexts (e.g., teacher and parent; Kraemer et al, 2002).

The unique perspective factors, P_{1t}, P_{2t},…, P_{Rt}, each affect only a single informant's ratings and are assumed to be orthogonal to C_t and to each other. By imposing the constraint that the factors are orthogonal, we ensure that each factor P captures variance that is unique to a specific informant and that is not shared (i.e., does not covary) with other informants (in contrast to C_t which represents shared variability across sets of ratings, including shared perspectives). Independent sources of variation across informants might include distinct contexts of observation, differences in the opportunity to observe behaviors, differences in informants' roles (e.g., mother, father, teacher), and subjective biases.

Finally, when the same item is rated by multiple informants, we anticipate that the item responses will be dependent not only due to the influence of the common factor but also due to the influence of a factor specific to that item. For instance, if the item "cries often" was administered as part of an assessment of negative affect, we would expect informant ratings on this item to reflect not just negative affect, globally defined across items, but also something specific to the behavior crying. The specific factors, S_{1t}, S_{2t},…, S_{It} account for this extra dependence. A given specific factor, S_{it}, is defined to affect the responses of all informants to item i but to no other items. The specific factors are also assumed to be orthogonal to one another and all other factors in the model. With these constraints, the specific factors capture covariation that is unique to a particular item. As noted by Marsh (1993), modeling specific factors for items rated by multiple informants is essential to avoid inflating the common factor variance.

Further restrictions on the model parameters should be imposed when some or all informants are interchangeable (Eid et al., 2008; Nussbeck et al., 2009). For example, suppose that a researcher has obtained self-ratings as well as the ratings from two randomly chosen peers for each target. The self-ratings are structurally different from the peer-ratings whereas the peers are interchangeable. The tri-factor model should be specified so that all parameters (e.g., item intercepts, factor loadings, and perspective factor means and variances) are

constrained to equality across interchangeable informants but allowed to differ across structurally different informants (Nussbeck et al., 2009).

For structurally different informants, the question may be raised whether *all* parameters should differ across informant types. For instance, mothers and fathers are structurally different, yet it may be that both parents engage in a similar process when rating their children. As in the broader literatures on measurement and structural invariance, similarity may be seen as a matter of degree, and this can be assessed empirically within the model through the imposition and testing of equality constraints (see, e.g., Gonzalez & Griffin, 1999; Kenny, Kashy & Cook, 2006, for tests of distinguishability among dyad members). Equal item intercepts and factor loadings would imply that informants interpret and respond to the items in the same way. If, additionally, perspective factor variances are equal then this would imply that the decomposition of variance in the item responses is identical across informants. Finally, if the perspective factor means are also equal, this would imply that there are no systematic differences across informants in their levels of endorsement of the items. Indeed, if all of these equality constraints can be imposed then the informants neither interpret the items differently nor does one type of informant provide systematically higher or more variable ratings than another, and the model obtains the same form as the model for interchangeable raters. In contrast, when some but not all equality constraints are tenable, this may elucidate important differences between informants. For instance, compared to mothers, fathers may be less likely to rate their children as displaying negative affect (Seiffge-Krenke & Kollmar, 1998), resulting in a mean difference between the unique perspective factors for mothers and fathers. As such, empirical tests of equality constraints across informants may provide substantively important information on whether and how the ratings of structurally different informants actually differ.

As with all latent variable models, some additional constraints are necessary to set the scale of the latent variables. We prefer to set the means and variances of the common and specific factors to zero and one, respectively. For interchangeable informants, we similarly standardize the scale of the perspective factors. In contrast, for structurally different informants, we standardize the scale of one perspective factor, while estimating the means and variances of the other perspective factors.[3] To set the scale of the remaining perspective factors, the intercept and factor loading for at least one item must be equated across informants. Last, when only two informants are present, the factor loadings for the specific factors must be equated across informants, in which case the specific factor essentially represents a residual covariance.

Standardizing the scale of the latent factors has the advantage that all non-zero factor loadings can be estimated and compared in terms of relative magnitude. Comparing $\lambda_{ir}^{(C)}$ to $\lambda_{ir}^{(P)}$ sheds light on the subjectivity of the item ratings. A high value for $\lambda_{ir}^{(P)}$ indicates that responses to this item largely reflect the idiosyncratic views of the informants, whereas a high value for $\lambda_{ir}^{(C)}$ indicates that the item responses largely reflect common opinions of the target's trait level. Similarly, if $\lambda_{ir}^{(S)}$ is large relative to $\lambda_{ir}^{(C)}$ and $\lambda_{ir}^{(P)}$ then this suggests that the item is not a particularly good indicator of the general construct of interest (as most of the variability in item responses is driven by the specific factor). Inspection of the relative magnitude of the factor loadings can thus aid in scale evaluation and development.

[3]Note that the model fit and interpretation is invariant to the choice of which perspective factor to standardize. This choice influences the means and variances of the remaining perspective factors, and hence also the scale of the perspective factor scores for these informants (though not the scale or scores of the common factor). Comparison of the perspective factor means, variances, and scores across informants are only meaningful if equality constraints can be imposed on the intercepts and factor loadings of all (or many) items, implying factorial invariance (or partial factorial invariance).

Conditional Tri-Factor Model

The conditional tri-factor model extends the model described above by including predictors of the different factors. Adding predictors to the model serves at least two potential purposes. First, by incorporating predictors we bring additional information into the model by which to improve our score estimates, a topic we will explore in greater detail in the next section. Second, we can evaluate specific hypotheses concerning sources of systematic variability in the factors. For instance, if we think that a given predictor influences trait levels, then we can regress the common factor on that predictor. Alternatively, if we think that informants with a particular background are more likely to rate a target's behavior in a specific direction, then we can use this background characteristic to predict the perspective factors. We might also regress perspective factors on contextual variables that vary within informant type, such as amount of time spent with the target. Finally, if we think that some items are more likely to be endorsed for certain targets than others (irrespective of their global trait levels) then we can regress the specific factors of these items on the relevant target characteristics.

It is conceptually useful to distinguish between predictors that vary only across targets versus predictors that vary across informants for a given target. For instance, if parental ratings of negative affect were collected, predictors of interest might include the child's gender, whether the mother has a lifetime history of depression and whether the father has a lifetime history of depression. Child gender is a target characteristic whereas history of depression is informant-specific. Designating target characteristics by the vector $\mathbf{w}_t$ and informant-specific characteristics by the vectors $\mathbf{x}_{rt}$ (one for each rater r), regression models for the factors can be specified as follows:

$$C_t = \alpha^{(C)} + \mathbf{w}_t' \boldsymbol{\beta}^{(C)} + \sum_{r=1}^{R} \mathbf{x}_{rt}' \boldsymbol{\gamma}_r^{(C)} + \zeta_t^{(C)} \quad (2)$$

$$P_{rt} = \alpha_r^{(P)} + \mathbf{x}_{rt}' \boldsymbol{\gamma}_r^{(P)} + \zeta_{rt}^{(P)} \quad (3)$$

$$S_{it} = \alpha_i^{(S)} + \mathbf{w}_t' \boldsymbol{\beta}_i^{(S)} + \zeta_{it}^{(S)} \quad (4)$$

where α designates an intercept term, $\boldsymbol{\beta}$ designates a vector of target effects, $\boldsymbol{\gamma}$ designates a vector of informant-specific effects, ζ designates unexplained variability in a factor, and parenthetical superscripts are again used to differentiate types of factors. The tri-factor model previously illustrated in Figure 1 is extended in Figure 2 to include target and informant-specific effects on the common factor and perspective factors. Specific factors can also be regressed on target characteristics but these paths are not included in Figure 2 to minimize clutter. For interchangeable informants, all parameters should again be equated across raters, including the effects contained in the vectors $\boldsymbol{\gamma}_r^{(C)}$ and $\boldsymbol{\gamma}_r^{(P)}$ in Equations (2) and (3), and the intercepts and residual variances of the perspective factors. For structurally different raters, these parameters may be permitted to differ, or may be tested for equality.

In Equation (2) the common factor is affected by both target and informant characteristics. Continuing our example, girls may be higher in negative affect than boys – a target effect. Having a depressed mother and/or father may also predict higher negative affect – an informant-specific effect. These effects on the common factor are shown as directed arrows in Figure 2 (child gender would be w and parental history of depression would be x_M and x_F for mothers and fathers, respectively). In contrast, for perspective factors, the focus is exclusively on informant-specific predictors. Equation (3) shows that the perspective factor

of a given informant is predicted by his or her own characteristics but not the characteristics of other informants. For instance, an informant with a history of depression might provide positively biased ratings of target negative affect. A history of depression for one parent does not, however, necessarily bias the ratings of the other parent. Thus, in Figure 2, x_M exerts an effect on the perspective factor for the mother and x_F exerts an effect on the perspective factor for the father. Last, the regression model for the specific factors includes only target effects. For instance, child gender could be allowed to influence the item "cries often" if this symptom was more likely to be expressed by girls than boys even when equating on levels of negative affect. The model shown in Figure 2 would then be extended to include a path from *w* to the specific factor for "cries often."

With the incorporation of these predictors, the assumptions previously made on the factors in the unconditional tri-factor model now shift to the residuals. Specifically, each residual ζ in Equations (2) to (4) is assumed to be normally distributed and orthogonal to all other residuals. Conditional normality and independence assumptions for the factors may be easier to justify in practice. For instance, suppose that parents with a lifetime history of depression provide positively biased ratings of their children's negative affect. This bias is detectable when one parent and not the other has experienced depression, but suppose that both parents have experienced depression. In the unconditional tri-factor model, this would result in a higher score for the child on the common factor for negative affect, because the bias is shared across both parents' ratings. The conditional tri-factor model, in contrast, would account for this shared source of bias, removing its contribution to the common factor and making the assumption of conditional independence for the perspective factors more reasonable.

Similarly, the constraints needed to identify the model shift from the factor means and variances in the unconditional model to the factor intercepts and residual variances in the conditional model. For the conditional model, we generally prefer to set the intercepts and residual variances of the common factor and specific factors to zero and one. For interchangeable raters, the same constraints are placed on the perspective factor intercepts and residual variances. In contrast, for structurally different raters, we prefer to standardize the scale of one perspective factor while freely estimating the intercepts and residual variances of the remaining factors. This choice requires the intercept and factor loading of at least one item to be set equal across informants (see Footnote 3).

These scaling choices are convenient but not equivalent to setting the marginal means and variances of the factors to zero and one, as we recommended for the unconditional model. Two implications of this scaling difference are noteworthy. First, the raw intercept and factor loading estimates for the items are not directly comparable between the unconditional and conditional models. Second, unlike the unconditional model in which the marginal variances of the factors were equated (except possibly for some perspective factors) the marginal variances of the factors will generally differ within the conditional model. For this reason, one cannot directly compare the magnitudes of the raw factor loading estimates across types of factors in the conditional model. In both instances, however, such comparisons can be made when using the conditional model by computing standardized estimates.

Potential Modifications to the Model Structure

In both the unconditional and conditional tri-factor model structures, we imposed certain assumptions to improve the clarity of the model and its interpretation. Not all of these assumptions are strictly necessary to identify the model and, in certain cases, some may be viewed as theoretically unrealistic or inconsistent with the data. These assumptions may then be relaxed, albeit with the risk of muddying the conceptual interpretation of the factors. We

revisit two of these assumptions here, beginning with the assumption that all factors are uncorrelated.

Within the unconditional model, the assumption of orthogonality enables us to define the perspective factors as the non-shared, unique components of variability in the item responses of the informants. It also allows us to state that the common and specific factors alone account for the shared variability across informants, with the common factor representing the broader construct of interest (as defined by the full item set) and the specific factors representing narrower constructs (as defined by single items). Within the conditional model, these same definitions of the factors pertain after controlling for the predictors (which may also account for some common variability across items and/or informants).

In some cases, however, a researcher may have a theoretical rationale for permitting correlations among a subset of factors in the model. For instance, if the content of two items overlapped, such as “often feels lonely” and “feels lonely even when with others,” then one might allow the specific factors for the two items to correlate. In effect, the correlation between these specific factors would account for the influence of a minor factor (loneliness) that jointly affects both items but that is narrower than the major factor of interest (negative affect). Failing to account for local dependence due to minor factors can lead to the locally dependent items “hijacking” the major factor, as evidenced by much higher factor loadings for these items relative to other items. For the tri-factor model, in particular, failing to account for local dependence among the specific factors would be expected to distort the factor loadings for the common factor, as this is the only other factor in the model that spans between informants. Introducing correlated specific factors may thus aid in avoiding model misspecifications that would otherwise adversely impact the estimation of the common factor. The trade off, however, is that the conceptual distinctions between the factors become blurred: the common factor no longer reflects all common variability across raters other than that unique to particular items, since some common variability is now accounted for by the correlations among specific factors. More cynically, the introduction of many correlated specific factors could be motivated solely by a desire to improve model fit, and might occlude a more profound misspecification of the model (e.g., a multidimensional common factor structure). The inclusion of correlated specifics should thus be justified conceptually (not only empirically).

As with the specific factors, the perspectives of some informants may be more similar than others. For instance, research on employee evaluations suggests that an average of 22% of the variance in job performance ratings can be attributed to common rater “source” effects (i.e., informants originating from the same context), over and above variance attributable to idiosyncratic rater characteristics (Hoffman, Lance, Bynum, & Gentry, 2010; see also Mount, Judge, Scullen, Sytzma, & Hezlett, 1998; Scullen, Mount, & Goff, 2000). Similarly, if the informants were mothers, fathers, and teachers, one might expect mother and father ratings to be more similar to one another than to teacher ratings. We might then find much higher factor loadings for the parental ratings than the teacher ratings, as the common factor would be required to account for the higher similarity of the parent ratings. Introducing a correlation between the parent perspective factors would account for overlapping role and context effects for mothers and fathers, enabling the common factor to integrate the across-role, across-context ratings of teachers and parents more equitably. It is important to recognize, however, that introducing correlated perspective factors for informants originating from a common setting or context changes the definition of the common factor, in effect reweighting how information is integrated across informants. The conceptual definition of the common factor is then determined predominantly by which perspective factors remain uncorrelated.

The other assumption that we shall revisit concerns which types of predictors are allowed to affect which types of factors. Equations (2) to (4) included only a subset of theoretically plausible associations between known characteristics of the targets and informants and the latent factors that underlie informant ratings. Specifically, we allowed both target and informant characteristics to affect the common factor, but we restricted the predictors of the perspective factors to informant characteristics and the predictors of the specific factors to target characteristics. These restrictions are not strictly necessary and could be relaxed if there was a conceptual motivation or an empirical imperative to do so. For instance, if mothers rate girls as higher than boys on negative affect but fathers do not, then the perspective factor for one or the other parent could be regressed on child gender by adding target characteristic effects to Equation (3). Cross-informant effects could potentially also be added to Equation (3). Similarly, Equation (4) could be expanded to include informant characteristics or item-specific predictors (e.g., a predictor differentiating whether an item is positively or negatively worded). We regard the simpler specifications in Equations (2) to (4) to be conceptually and practically useful for a wide variety of potential applications of the tri-factor model, but decisions about which predictors to include should be driven by the theoretical underpinnings of a given application.

Aside from these two assumptions of the model, we shall also reconsider one assumption we have made concerning the data. To this point, we have assumed for simplicity that item sets are parallel across informants, i.e., that all informants provide ratings on the same set of items. This parallelism is theoretically desirable because it allows for the estimation of the specific factors and the separation of specific factor variance from random error. But the use of parallel items may not always be feasible, particularly when informants observe the target in different contexts with different affordances for behavior (e.g., behaviors at school or work versus behaviors at home). The tri-factor model is still applicable with non-parallel item sets; however, for items rated by only one informant the specific factors must be omitted from the model specification. For these items, the analyst must be mindful that specific factor variance is conflated with random error. One implication is that the reliability of the item may be under-estimated. Nevertheless, perspective and common factor loadings can be estimated and compared for these items. For structurally different informants, equivalence tests should be restricted to the subset of parallel items (if any).

Estimation

Estimation of the tri-factor model is straightforward if all ratings are made on a continuous scale. The model can then be fit by maximum likelihood in any of a variety of structural equation modeling (SEM) software programs (or by other standard linear SEM estimators). When items are binary or ordinal, however, estimation is more challenging. We provide a cursory review of this issue here; for a more extensive discussion of estimation in item-level factor analysis see Wirth and Edwards (2007).

Maximum likelihood estimation with binary or ordinal items is complicated by the fact that the marginal likelihood contains an integral that does not resolve analytically and must instead be approximated numerically. Unfortunately, the computing time associated with the most common method of numerical approximation, quadrature, increases exponentially with the number of dimensions of integration. As typically implemented, the dimensions of integration equal the number of factors in the model. Given the large number of factors in the tri-factor model ($1 + R + I$), this approach would seem to be computationally infeasible. Cai (2010a), however, noted that for certain types of factor analytic models, termed two-tier models, the dimensions of integration can be reduced. The tri-factor model can be viewed as a two-tier model where the common and perspective factors make up the first tier and the specific factors make up the second tier. Using Cai's (2010a) approach, the dimensions of

integration for the tri-factor model can then be reduced to a more manageable number and computationally efficient estimates can be obtained by maximum likelihood using quadrature. Another option is to implement a different method of numerical approximation, for instance Monte Carlo methods. In particular, the Robbins-Monro Metropolis-Hastings maximum likelihood algorithm developed by Cai (2010b, 2010c) is also a computationally efficient method for fitting the tri-factor model. Similarly, Bayesian estimation by Markov Chain Monte Carlo (MCMC) approximates maximum likelihood when priors are selected to be non-informative (Edwards, 2010).

An alternative way to fit the tri-factor model is to use a traditional, limited-information method of estimation that is somewhat less optimal statistically but quite efficient computationally. Motivating this approach is the notion that binary and ordinal responses can be viewed as coarsened versions of underlying continuous variables (Christoffersson, 1975; Muthén, 1978; Olsson, 1979). For instance, the binary item “lonely” reflects an underlying continuous variable “loneliness.” Assuming the underlying continuous variable to be normally distributed corresponds to a probit model formulation (e.g., choosing g in Equation (1) to be the probit link function). Based on this assumption and the observed bivariate item response frequencies, polychoric correlations can be estimated for the underlying continuous variables. Finally, the model is fit to the polychoric correlation matrix using a weighted least squares estimator. The theoretically optimal weight matrix is the asymptotic covariance matrix of the polychoric correlations but this weight matrix is unstable except in extremely large samples (Browne, 1984; Muthén, du Toit & Spisic, 1997). In practice, a diagonal weight matrix is often employed as a more stable alternative (Muthén et al, 1997). Simulation studies have demonstrated that this Diagonally Weighted Least Squares (DWLS) estimator performs well even at relatively modest sample sizes (Flora & Curran, 2004; Nussbeck et al., 2006).

Several issues beyond computational efficiency may also influence estimator selection. Practically, weighted least squares and DWLS are widely accessible and easily implemented in a variety of SEM software programs. Further, analysts using these estimators have access to well-developed tests of model fit and goodness of fit criteria from which to judge the suitability of the model for the data. ML and Bayesian MCMC estimation approaches, however, more favorably accommodate missing data. In this context, missing data is most likely to occur due to informant non-response, such that all items for a given informant are missing. Both ML and MCMC estimation include cases with partial data under the assumption that the missing data are missing at random. In contrast, weighted least squares estimators are often implemented under the assumption of complete data, requiring listwise deletion and implicitly assuming missing data are missing completely at random. Within some software (e.g., Mplus), partially missing data is permitted with weighted least squares or DWLS under the assumption that the missing data process may be covariate-dependent (Asparouhov & Muthén, 2010).[4]

Another issue that may influence estimation is factor loading reflection (Loken, 2005). Specifically, an equivalent fit to the data can be obtained by reversing the polarity of a factor (e.g., multiplying all factor loadings for a factor by negative one). In practice this problem is most likely to affect doublet specific factors (occurring when only two informants rate each item). For maximum likelihood and weighted least squares estimation factor loading reflection is largely a nuisance that can be avoided by providing positive start values for the factor loadings or by judiciously implementing positivity constraints. With MCMC estimation, however, factor loading reflection can be a more serious problem as the iterative process constructs a bimodal posterior distribution by sampling positive and negative values

[4]See Schafer and Graham (2002) for an accessible overview of missing data processes and their implications for data analysis.

for the factor loadings, but "converging" on neither estimate. Informative priors, particularly for doublet specific factor loadings, can help to mitigate this problem.

Scoring

After fitting the model it will often be of interest to obtain score estimates for the sample that can be used in later data analyses. For instance, the primary goal behind our development of the tri-factor model was to obtain negative affect scores that we could subsequently use in longitudinal analyses to predict the onset of substance use disorders (Hussong et al., 2011). Rather than include the mother's ratings and father's ratings of the child as separate measures, we wished to obtain a single, integrated, multi-informant measure that would be purged of the idiosyncratic views of the specific informants, including potential rater bias. Thus our focus was on obtaining valid and reliable common factor scores. In other applications, however, the perspective factor score estimates may be of equal or greater interest. For instance, Lance et al. (2008) and De Los Reyes (2011) argue that rating discrepancies across informants are substantively meaningful and should not be regarded simply as a type of measurement error.

Score estimates for the factors are obtained similarly regardless of the method of estimation used to fit the tri-factor model. Specifically, score estimates are computed as either the mean or the mode of the posterior distribution of the factor for the individual given his or her observed item responses and the parameter estimates for the model (see Skrondal & Rabe-Hesketh, 2004, sections 7.2–7.4). With continuous items, the mean and mode coincide and can be computed via the regression method (Bartholomew & Knott, 1999; Thomson, 1936, 1951; Thurstone, 1935). With binary or ordinal items the mean and mode are not equal but tend to be highly correlated. In the literature on item response theory the mean is usually referred to as the Expected a Posteriori (EAP) estimate and the mode is usually referred to as the Modal a Posteriori (MAP) estimate, with the latter being somewhat easier to compute (Thissen & Orlando, 2001).[5]

Both EAPs and MAPs are "shrunken" estimates, meaning that the scores generated for the target will be closer to the factor mean (across targets) as the amount of information available for the target (e.g., number of items rated) decreases. Conceptually, we are using what we know about the population in general to improve our score estimates for each specific individual. In the unconditional tri-factor model, all scores for a given factor are shrunken toward the same marginal mean. In contrast, in the conditional tri-factor model, scores are shrunken toward the conditional mean of the factor given the values of the predictors (Bauer & Hussong, 2009). In other words, rather than use the overall average to improve our score estimate for the target, we can use the average for people who are similar to the target with respect to the predictors. For example, if the common factor is regressed on the sex of the target, then scores for girls will be shrunken toward the conditional mean for girls and scores for boys will be shrunken toward the conditional mean for boys. In this sense, the scores obtained from a conditional tri-factor analysis are "tuned" to the characteristics of the target and informants.

Summary

In total, the tri-factor model provides a number of key advantages for modeling multi-informant data. First, the model does not require ratings on multiple traits or multiple targets, as the focus is not on construct validity but on construct measurement. Second,

[5]Note that even if the factors are restricted to be marginally orthogonal in the fitted model the score estimates may nevertheless be correlated. For continuous items, one can ensure orthogonality of the score estimates by using the alternative scoring method of Anderson and Rubin (1956). The regression method score estimates are, however, more efficient (having smaller standard errors).

because the model is fit to item-level data from multiple informants, it is possible to evaluate item quality in a way that is not possible when analyzing scale-level data or single-informant item-level data. Third, the conditional formulation of the tri-factor model permits tests of hypotheses about putative sources of trait variability, informant differences, and item properties. Finally, the model can be used to create and evaluate scores for the factors for use in subsequent analyses. When generated from the conditional tri-factor model these scores are tuned to the specific characteristics of the targets and informants.

Let us now turn to an empirical application of the tri-factor model to illustrate these advantages of the model.

Example: Parent-Reported Negative Affect

Our demonstration derives from an integrative data analysis of two longitudinal studies of children of alcoholic parents and matched controls (children of non-alcoholic parents): the Michigan Longitudinal Study (MLS; Zucker et al., 1996, 2000) and the Adolescent and Family Development Project (AFDP; Chassin, Rogosch & Barrera, 1991). As noted by Curran et al. (2008), a major challenge in conducting integrative data analysis is measurement (see also Bauer & Hussong, 2009; Curran & Hussong, 2009). In combining longitudinal studies, in particular, one must be sensitive to age-related changes in the construct and the age-appropriateness of the items. In the present case, we sought to obtain a measure of negative affect for children between 2 and 18 years of age based on ratings provided by both mothers and fathers. In fitting the tri-factor model to this data our aims were threefold. First, we sought to explicate the sources of variance, both random and systematic, that underlie parent ratings of negative affect. Second, we wished to evaluate potential study, age, familial risk, and gender differences in negative affect, both as broadly defined across the item set and as narrowly measured by specific items. Third, we sought to generate valid and reliable negative affect scores that would account for potential rater biases, for instance due to parental depression.

Sample

Like many psychometric models, the tri-factor model assumes independence of observations. This assumption would be violated if we applied the model directly to the full set of longitudinal data. We thus pursued the strategy recommended by Curran et al. (2008) to select ratings randomly from a single age for each participant for inclusion in the tri-factor analysis. We refer to this cross-sectional draw from the data as the *calibration sample* (N=1080). It is this sample that is used to fit, evaluate, and refine the model. Once the optimal model has been determined, however, the estimates obtained from the calibration sample can be used to generate factor score estimates for the full set of observations, facilitating subsequent longitudinal analyses.

To check the stability of our results, we also randomly selected a second set of ratings for each target (excluding the ages selected for the calibration sample) and refit the final model. We shall refer to this second cross-sectional sample as the *cross-validation sample* (N=975). More ideally, we would cross-validate the model on a truly independent sample; nevertheless, this second sample provided an opportunity to evaluate the stability of the parameter estimates and scores obtained from the model.

Table 1 shows the number of observations at each age from each study present in the original longitudinal sample, the calibration sample, and the cross-validation sample.

Measures

Thirteen binary items present in both the MLS and AFDP studies (originating from the Child Behavior Checklist; Achenbach & Edelbrock, 1981) were identified as indicators of negative affect for inclusion in the tri-factor analysis, as shown in Table 2. For conditional models, target-specific characteristics of interest were study (58% from MLS), gender (64% male), and age (range 2–18; *M*=12.07, *SD*=3.84). Informant-specific characteristics of interest were lifetime history of an alcohol use disorder (AUD; 24% of mothers, 65% of fathers), depression or dysthymia (14% of mothers, 9% of fathers), or antisocial personality disorder (ASP; 1% of mothers, 12% of fathers). ASP almost always co-occurred with an AUD, thus impairment was assessed via three binary parental impairment variables indicating: (1) history of depression or dysthymia, (2) history of an AUD without ASP, and (3) history of an AUD with ASP.

Fitting the tri-factor model

Unconditional model—We fit the model shown in Figure 1 to the calibration sample data using the WLSMV (DWLS) estimator in M*plus* version 6.1 (Muthén & Muthén, 1998–2010).[6] Because mothers and fathers are structurally different informants, the underlying process by which they rate the negative affect of their child may differ. As such, we initially allowed the model parameters to differ across the two informants. Each factor was scaled to have a mean of zero and variance of one with the exception that the father perspective factor mean and variance were freely estimated. Identification constraints were imposed to equate the intercepts and factor loadings of one item across mothers and fathers. Since the specific factors were doublets, we avoided the factor loading reflection problem by imposing boundary constraints on the loadings of these factors. Overall, this model provided good fit to the data, $\chi^2(261) = 550.69$, $p < .0001$; RMSEA = .03 (90% CI = .028–.036); CFI = .96; TLI = .95.

We next evaluated the degree of structural similarity between mothers' and fathers' ratings. We began by imposing equality constraints only on the item intercepts and factor loadings (i.e., factorial invariance). The fit of this model was still good, $\chi^2(297) = 569.73$, $p < 0001$; RMSEA = .03 (90% CI = .026–.033); CFI = .96; TLI = .96, but significantly worse than the unrestricted model, $\Delta\chi^2(36)= 62.50$, p=.004.[7] Because the chi-square difference test is sensitive to sample size, potentially having power to detect even trivial differences between parameter values, Cheung and Rensvold (2002) suggested retaining invariance constraints that do not lead to a meaningful decrement in goodness of fit indices (i.e., RMSEA, CFI, TLI). More recently, however, Fan and Sivo (2009) argued that changes in goodness of fit indices are insensitive to misspecification of the mean structure, particularly in large models. Consistent with the latter observation, further inspection of the results suggested that, all else being equal, fathers were more likely to endorse Item 3, and mothers were more likely to endorse Items 5, 6, and 12 (see Table 2). Allowing only the intercepts of these four items to differ across informants resulted in a non-significant chi-square difference relative to the unrestricted model, $\Delta\chi^2(32)= 45.02$, p=.06.

Given this partial invariance of the factor structure across mothers and fathers, we proceeded to test whether the perspective factor means and variances differed between the two

[6]We also fit the unconditional tri-factor model using the two-tier and Robbins-Monro maximum likelihood algorithms as well as by MCMC. Accounting for differences between the logit and probit scales, the parameter estimates obtained by these methods were similar to those obtained by DWLS and the factor score estimates were very highly correlated.

[7]Use of the WLSMV estimator within Mplus produces a robust, mean- and variance-adjusted chi-square test statistic of overall model fit. Chi-square difference tests between nested models cannot be computed as the simple difference in these values. Therefore, throughout this manuscript, robust chi-square difference tests were computed using the approach developed by Satorra (2000) and Satorra and Benlter (1999) and implemented in Mplus with the DIFFTEST command (Asparouhov & Muthén, 2006).

informants. Equating these parameters (i.e., setting the perspective factor mean and variance to zero and one, respectively, for both informants) did not significantly worsen the fit of the model, $\Delta\chi^2(2) = 1.34$, p=.51, and the absolute fit of this model was also good, $\chi^2(295) = 511.16$, $p < 0001$; RMSEA = .03 (90% CI = .022–.030); CFI = .97; TLI = .97. Thus, mothers and fathers functioned similarly to interchangeable raters, with the important exception that they differentially endorsed four out of thirteen items.

Conditional model—In extending to the conditional tri-factor model we adopted the scaling convention to set the intercept and residual variance of the factors to zero and one, respectively. The model was then fit in a sequence of steps, ordered by theoretical priority. First, we regressed the common factor on potential sources of target variability. We evaluated target effects of study, gender, age (including linear, quadratic and cubic trends), and all two-way interactions between these predictors, trimming non-significant interactions from the final model.[8] Referencing Figure 2, this block of predictors replaces the single predictor *w*. Simultaneously, we included informant-specific effects of parental impairment on the common factor.[9] The parental impairment variables thus replace x_M and x_F in Figure 2, with paths included from these predictors to the common factor. The effects of the parental impairment variables were initially allowed to differ over informants, and this model provided good fit to the data, $\chi^2(596) = 876.41$, $p < 0001$; RMSEA = .02 (90% CI = .019–.025); CFI = .95; TLI = .95. Constraining these effects to be equal did not significantly worsen model fit, $\Delta\chi^2(3) = 2.10$, p=.55; $\chi^2(599) = 860.60$, $p < 0001$; RMSEA = .02 (90% CI = .018–.025); CFI = .96; TLI = .95. We thus retained these equality constraints.

Second, we regressed each perspective factor on the impairment indicators for the corresponding informant. Referring again to Figure 2, we added the paths from x_M and x_F to P_M and P_F, where each *x* stands for a set of impairment variables. Here it is important to differentiate the two effects of the parental impairment variables. The regression of the common factor on the impairment variables captures the potentially real elevation of negative affect of children with one or more impaired parents. To the extent that the negative affect of children of impaired parents is actually higher than that of other children, this should be reflected in the ratings of both parents, as transmitted by the common factor, irrespective of which parent might be impaired. Prior literature suggests, however, that impaired parents may also provide artificially elevated ratings of negative affect. The regression of the perspective factors on the impairment variables captures this possible source of rater bias, which should be observed only in the ratings of the impaired parent and not an unimpaired co-parent. We again initially allowed these effects to differ by informant, $\chi^2(593) = 846.54$, $p < 0001$; RMSEA = .02 (90% CI = .018–.024); CFI = .96; TLI = .95, and then constrained them to be equal, $\chi^2(596) = 838.79$, $p < 0001$; RMSEA = .02 (90% CI = .017–.024); CFI = .96; TLI = .96. As the chi-square difference test ($\Delta\chi^2(3) = 2.37$, p=.50) was not significant, we retained the more parsimonious structure with equal effects across informants.

The fit of the conditional tri-factor model at this step was already quite good; nevertheless, to be conservative, we proceeded to evaluate potential target effects on the specific factors that, if falsely excluded from the model, might distort the pattern of effects observed for the common factor (akin to the assessment of differential item functioning in item response theory). Because we had no theoretical predictions concerning the specific factors, we identified these effects using an empirical approach. Specifically, we examined modification indices to identify items for which target-specific effects might explain systematic variation. The mechanical use of modification indices in structural equation models has been criticized

[8]To facilitate estimation and interpretation given the presence of power terms and interactions, age was centered at 10 years.
[9]Given some missing data on the impairment variables, the sample size for this and subsequent models was reduced to N=947.

(appropriately) for failing to identify model misspecifications accurately (MacCallum, Roznowski and Necowitz, 1992), but more targeted uses of modification indices in measurement models have proven beneficial (Glas, 1998; Yoon & Millsap, 2007). Additionally, to reduce the likelihood of capitalizing on chance, we adopted a conservative criterion for effect inclusion: a modification index exceeding 6.64, the critical value for a single degree-of-freedom chi-square test with an alpha level of .01. Predictors of the specific factors were added to the model one at a time, beginning with the effect displaying the largest modification index. Using this approach we detected target effects for the specific factors of Items 1, 2, 3, 4, 8, 9 and 12 (see Table 2). The fit of the final model including these effects was excellent, $\chi^2(586) = 718.34$, $p = .0001$; RMSEA = .015 (90% CI = .011–.019); CFI = .98; TLI = .98.

Interpretation

Raw and standardized intercept and factor loading estimates for the final tri-factor model are presented in Table 3. The standardized solution is particularly informative as the magnitudes of the standardized factor loadings are directly comparable and indicate the relative effects of the common, perspective, and specific factors on the items. Comparing columns of Table 3, we can see that the common factor loadings are often lower than the perspective and specific factor loadings. Thus the negative affect common factor often contributes less to the item ratings than variation uniquely associated with the informant or that is specific to a given item. Comparing rows of Table 3, we can see which items most reflect the common factor and are least susceptible to perspective differences. At the extreme, we can see that the item “has to be perfect” has a very small loading on the common factor. Endorsement of this item co-occurs with other items almost exclusively due to unique perspective effects, something that would not be revealed in a factor analysis of a single informant’s ratings. It is encouraging to note, however, that the items that might be considered core features of negative affect also tend to have the highest common factor loadings (e.g., Items 1, 5, 6, 10, 11, and 13).

Table 4 presents raw and partially standardized estimates for the effects of the regressors on the factors. The partially standardized estimates are computed by standardizing the factors but leaving the regressors in their raw scales, and are particularly useful when predictors are binary (e.g., study, gender, and impairment) or have a meaningful metric (e.g., age). Sex-specific patterns of developmental change were detected for the common factor, as depicted in Figure 3. A study difference was also observed for the common factor: targets from MLS displayed lower levels of negative affect than targets from AFDP. In addition, parental impairment effects on the common and perspective factors indicate that depressed or ASP parents indeed have children with higher negative affect levels, but depressed parents in particular perceive the negative affect of their children to be even greater than it is commonly perceived to be (i.e., their ratings are biased).

The last effects listed in Table 4 are for the specific factors. Most of these effects are for study, with some items being endorsed less frequently in the MLS sample than would be expected due to study differences on the common factor alone. Age trends are observed for several items as well. In particular, “lonely” and “cries a lot” show steeper declines with age than the common factor. “Has to be perfect” also follows a distinct age trend, but this is not terribly surprising given that this item did not load on the common factor. Finally, there is a gender difference on the item “cries a lot.” This item is more likely to be endorsed for girls than boys even after accounting for gender differences in the common factor of negative affect.

Sensitivity Analysis

To examine the stability of the final model we refit the final conditional tri-factor model to the cross-validation sample, once again obtaining excellent fit, χ^2 (586) = 782.42, $p < .0001$; RMSEA = .02 (90% CI = .016–.023); CFI = .97; TLI = .96. The intercepts, loadings, and factor regression parameter estimates obtained from the two samples are compared in Figure 4. Although some discrepancies were observed for more extreme estimates, the two sets of estimates were generally quite similar and were correlated at .91. We also examined the stability of the scores obtained from the final model. Specifically, we generated both common factor and perspective MAP factor scores for the full longitudinal data based on each set of estimates. The correlations between the two sets of scores were between .97 and .98, as shown in the scatterplots in Figure 5, indicating a high level of stability.

Comparison to Usual Practice

We also considered whether the scores generated by the tri-factor model differed meaningfully from what might be obtained using more conventional strategies. The alternative scoring strategies we implemented ranged from simple to complex. The simplest approach was to average the proportion of items endorsed by the mother with the proportion of items endorsed by the father. The next simplest approach was to compute the proportion of items endorsed by *either* the mother *or* the father. The more complex approaches we considered both involved factor analyzing the item responses of mothers and fathers separately, obtaining factor scores for each reporter and then an average factor score. In the first variant of this approach we obtained the factor scores from a standard two-parameter logistic item response theory model (without differential item functioning). In the second variant, we implemented a moderated nonlinear factor analysis model to allow for predictor effects on the factor as well as potential differential item functioning (Bauer & Hussong, 2009). In contrast to the tri-factor model, the appropriate application of these alternative scoring approaches is less clear when data is missing for an informant. In those instances scores were based solely on the ratings of the available informant.

The relations among the scores, as observed within the calibration sample, are summarized in correlation form in Table 5. Squaring the correlations, we can see that the common factor scores obtained from the tri-factor model share approximately 60% to 70% of their variance with the scores obtained from more conventional approaches. By contrast, it is interesting that the conventional scores are much more highly correlated, sharing between 86% and 94% of their variance, despite considerable differences in the complexity of the methods by which they were obtained.

It is clear from these results that the common factor scores generated from the tri-factor analysis overlap with but are distinct from the scores provided by other approaches. The shared variance between the scores reflects the fact that each approach taps into the consensus view of the target to some extent. We would argue, however, that the tri-factor model provides a more interpretable measure of the target because it formally separates the unique perspective factor variance from the common factor variance. With ad hoc scoring approaches perspective effects (including rater bias) will often contaminate the combined scores, particularly when there are relatively few informants and these effects are unlikely to "average out." Indeed, when we examined the correlations between the ad hoc measures and the perspective factor scores obtained from the tri-factor analysis, we found these correlations to be substantial (r = .47–.54).

Conclusions

Relative to simpler approaches for aggregating multiple informant data, psychometric models offer an important advantage: they embody an explicit theoretical model of the sources of variation in informant ratings. The tri-factor model proposed here stipulates that the observed item responses reflect three sources of variation: a common factor, a perspective factor for the informant, and a specific factor for the item. In contrast to models developed for MTMM data, which largely focus on issues of construct validation, the tri-factor model is principally intended to serve as a measurement model. The tri-factor model provides information on the quality of the items; that is, the extent to which the responses to each item reflect the common factor versus perspective factors or specific factors. Additionally, predictors may be incorporated into the model to test putative sources of underlying target differences and/or rater bias. Finally, we can obtain scores from the tri-factor model that separate the common view of the target from the unique views of the individual informants, and that reflect the specific characteristics of the target and informants.

Our development of the tri-factor model draws upon a 100-year tradition in factor analysis of defining and interpreting factors in terms of the patterns of dependence they induce among the manifest indicators (beginning with Spearman, 1904). The tri-factor model also represents a straightforward extension of the bi-factor model originally developed by Holzinger and Swineford (1937), with a novel application to multiple informant data (for other recent extensions and applications of the bi-factor model within psychological research see Cai, Yung and Hansen, 2011; Gibbons and Hedecker, 1992; Gibbons et al, 2007; Reininghaus et al, 2011; Reise et al, 2011). In this sense, the tri-factor model is fully consistent with factor analytic theory. It is important to recognize, however, that the conventional approach to specifying and interpreting factor analytic models has sometimes been criticized. Two specific concerns are that model restrictions (e.g., independence of factors) sometimes appear arbitrary and that the meaning of the latent variables is inferred from intuition rather than formal psychometric theory (e.g., true score/classical test theory). These issues are discussed in detail for MTMM model specifications in Eid (2000), Pohl and Steyer (2010) and Pohl, Steyer & Kraus (2008), and with respect to closely related latent state-trait models in Steyer (1989), Steyer, Ferring and Schmitt (1992), and Geiser and Lockhart, (2012). The tri-factor model is potentially open to similar critiques. Yet we are of the opinion that there are many different and useful ways to define and interpret latent variables (see Bollen, 2002) and that a broad, inclusive approach enables the development of models that are both conceptually appealing and practically useful. In this sense, the ultimate utility of the tri-factor model will be borne out through empirical applications, similar to our analysis of parent-reported negative affect.

There are, of course, many potential directions for future research on the tri-factor model. For instance, the data analytic conditions under which the tri-factor model can be usefully applied are presently unclear. Simulations examining parameter recovery across variations in sample size, number of items, item type (continuous versus binary or ordinal), number of informants, extent of missing data, number of predictors, and methods of estimation would be informative in this regard, but lie outside of the scope of the present study. The stability of our empirical results and the face validity of our findings are suggestive that the model will generally perform well with reasonably large samples. Another interesting avenue of research would be to apply the model in an experimental setting in which some informants are intentionally manipulated to provide biased ratings, to determine the extent to which the model can adequately detect this bias and remove it from the common factor score estimates.

Additionally, because the tri-factor model is a type of SEM, there are many potential ways that the model could be extended. For instance, one could include latent predictors to estimate regressor effects without biases due to measurement error. Alternatively, one could include criterion variables predicted by the common factor or perspective factors. Given the challenges of estimating these models, however, we believe that the greatest utility of the tri-factor model is as a measurement model for evaluating item quality and for generating scores. In most cases an investigator will wish to generate scores for the common factor, and perhaps also for perspective factors, for use in subsequent analyses. For instance, the motivation behind our example analysis was to generate negative affect scores so that we could examine whether trajectories of negative affect over childhood and adolescence could be used to predict substance use disorders in adulthood. The tri-factor model provided us with a way to generate an integrated negative affect measure based on ratings from both the mother and father while also controlling for possible rater biases. It is our hope that other investigators will similarly find the tri-factor model useful for understanding and summarizing the information provided by multiple informants in their own research.

Acknowledgments

We would like to thank Alison Burns for her assistance with this work. The project described was supported by Award Number R01DA015398 (Principal Investigators: Andrea Hussong and Patrick Curran) from the National Institute on Drug Abuse and Award Numbers R01AA016213 (Principal Investigator: Laurie Chassin) and R37AA07065 (Principal Investigator: Robert Zucker) from the National Institute on Alcohol Abuse and Alcoholism.

References

Achenbach TM. Commentary: Definitely more than measurement error: but how should we understand and deal with informant discrepancies? Journal of Clinical Child & Adolescent Psychology. 2011; 40:80–86.10.1080/15374416.2011.533416 [PubMed: 21229445]

Achenbach T, Edelbrock C. The classification of child psychopathology: A review and analysis of empirical efforts. Psychological Bulletin. 1981; 85:1275–1301.10.1037//0033-2909.85.6.1275 [PubMed: 366649]

Achenbach TM, Krukowski RA, Dumenci L, Ivanova MY. Assessment of adult psychopathology: meta-analyses and implications of cross-informant correlations. Psychological Bulletin. 2005; 131:361–382.10.1037/0033-2909.131.3.361 [PubMed: 15869333]

Achenbach TM, McConaughy SH, Howell CT. Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. Psychological Bulletin. 1987; 101:213–232.10.1037/0033-2909.101.2.213 [PubMed: 3562706]

Alessandri G, Vecchione M, Tisak J, Barbaranelli C. Investigating the nature of method factors through multiple informants: Evidence for a specific factor? Multivariate Behavioral Research. 2011; 46:625–642.10.1080/00273171.2011.589272

Anderson, TW.; Rubin, H. Statistical inference in factor analysis. In: Neyman, J., editor. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1956. p. 111-150.

Asparouhov, T.; Muthen, B. Robust chi square difference testing with mean and variance adjusted test statistics. Mplus web notes: No. 10. 2006 May 26. Retrieved from *Mplus* software website: http://statmodel.com/download/webnotes/webnote10.pdf

Asparouhov, T.; Muthén, B. Weighted least squares estimation with missing data. 2010 Aug 14. (Technical Appendix). Retrieved from M*plus* software website: http://www.statmodel.com/download/GstrucMissingRevision.pdf

Barbaranelli C, Fida R, Paciello M, Di Giunta L, Caprara GV. Assessing personality in early adoelscence through self-report and other-ratings in a multitrait-multimethod analysis of the BFQ-C. Personality and Individual Differences. 2008; 44:876–886.10.1016/j.paid.2007.10.014

Bartholomew, DJ.; Knott, M. Latent variable models and factor analysis. 2. London: Arnold; 1999.

Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. Psychological Methods. 2009; 14:101–125.10.1037/a0015583 [PubMed: 19485624]

Biesanz JC, West SG. Towards understanding assessments of the Big Five: multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. Journal of Personality. 2004; 72:845–876.10.1111/j.0022-3506.2004.00282.x [PubMed: 15210019]

Bollen KA. Latent variables in psychology and the social sciences. Annual Review of Psychology. 2002; 53:605–634.10.1146/annurev.psych.53.100901.135239

Bollen KA, Paxton P. Detection and determinants of bias in subjective measures. American Sociological Review. 1998; 63:465–478.10.2307/2657559

Boyle MH, Pickles AR. Influence of maternal depressive symptoms on ratings of childhood behavior. Journal of Abnormal Child Psychology. 1997; 25:399–412.10.1023/A:1025737124888 [PubMed: 9421748]

Browne MW. Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology. 1984; 37:62–83.10.1111/j.2044-8317.1984.tb00789.x [PubMed: 6733054]

Bullock BM, Deater-Deckard K, Leve LD. Deviant peer affiliation and problem behavior: A test of genetic and environmental influences. Journal of Abnormal Child Psychology. 2006; 34:29–41.10.1007/s10802-005-9004-9 [PubMed: 16550453]

Cai L. A two-tier full-information item factor analysis model with applications. Psychometrika. 2010a; 75:581–612.10.1007/s11336-010-9178-0

Cai L. High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. Psychometrika. 2010b; 75:33–57.10.1007/s11336-009-9136-x

Cai L. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. Journal of Educational and Behavioral Statistics. 2010c; 35:307–335.10.3102/1076998609353115

Cai L, Yung JS, Hansen M. Generalized full-information item bifactor analysis. Psychological Methods. 2011; 16:221–248.10.1037/a0023350 [PubMed: 21534682]

Campbell DT, Fiske DW. Convergent and discriminant validation by the mutlitrait-multimethod matrix. Psychological Bulletin. 1959; 56:81–105.10.1037/h0046016 [PubMed: 13634291]

Chassin L, Rogosch F, Barrera M. Substance use and symptomatology among adolescent children of alcoholics. Journal of Abnormal Psychology. 1991; 100:449–463.10.1037/0021-843X.100.4.449 [PubMed: 1757658]

Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling. 2002; 9:233–255.10.1007/BF02291477

Christoffersson A. Factor analysis of dichotomized variables. Psychometrika. 1975; 40:5–32.10.1007/BF02291477

Conway JM, Huffcutt AI. Psychometric properties of multisource performance ratings: a meta-analysis of subordinate, supervisor, peer, and self-ratings. Human Performance. 1997; 11:29–55.10.1207/s15328007sem1104_3

Conway JM, Lievens F, Scullen SE, Lance CE. Bias in the correlated uniqueness model for MTMM data. Structural Equation Modeling. 2004; 11:535–559.10.1207/s15327043hup1004_2

Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA. Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. Developmental Psychology. 2008; 44:365–380.10.1037/0012-1649.44.2.365 [PubMed: 18331129]

Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100.10.1037/a0015914 [PubMed: 19485623]

De Los Reyes A. 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. 2011; 40:1–9.10.1080/15374416.2011.533405 [PubMed: 21229439]

Edwards MC. A Markov Chain Monte Carlo approach to confirmatory item factor analysis. Psychometrika. 2010; 75:474–497.10.1007/s11336-010-9161-9

Eid M. A multitrait-multimethod model with minimal assumptions. Psychometrika. 2000; 65:241–261.10.1007/BF02294377

Eid M, Lischetzke T, Nussbeck FW, Trierweiler LI. Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple-indicator CT-C(*M*-1) Model. Psychological Methods. 2003; 8:38–60.10.1037/1082-989X.8.1.38 [PubMed: 12741672]

Eid M, Nussbeck FW, Geiser C, Cole DA, Gollwitzer M, Lischetzke T. Structural equation modeling of multitrait-multimethod data: different models for different types of methods. Psychological Methods. 2008; 13:230–253.10.1037/a0013219 [PubMed: 18778153]

Fan X, Sivo SA. Using Δgoodness-of-fit indexes in assessing mean structure invariance. Structural Equation Modeling. 2009; 16:54–69.10.1080/10705510802561311

Fergusson DM, Lunskey MT, Horwood LJ. The effect of maternal depression on maternal ratings of child behavior. Journal of Abnormal Child Psychology. 1993; 21:245–269.10.1007/BF00917534 [PubMed: 8335763]

Flora DB, Curran PJ. An evaluation of alternative methods for confirmatory factor analysis with ordinal data. Psychological Methods. 2004; 9:466–491.10.1037/1082-989X.9.4.466 [PubMed: 15598100]

Geiser C, Lockhart G. A comparison of four approaches to account for method effects in latent state-trait analyses. Psychological Methods. 2012; 17:255–283.10.1037/a0026977 [PubMed: 22309958]

Gibbons RD, Hedeker DR. Full-information item bi-factor analysis. Psychometrika. 1992; 57:423–436.10.1007/BF02295430

Gibbons RD, Bock RD, Hedeker D, Wiess DJ, Segawa E, Bhaumik DK, Kupfer DJ, Frank E, Grochocinski VJ, Stover A. Full-information item bifactor analysis of graded response data. Applied Psychological Measurement. 2007; 31:4–19.10.1177/0146621606289485

Glas CAW. Detection of differential item functioning using Lagrange multiplier tests. Statistica Sinica. 1998; 8:647–667.

Gonzalez R, Griffin DW. The correlation analysis of dyad-level data in the distinguishable case. Personal Relationships. 1999; 6:449–469.10.1111/j.1475-6811.1999.tb00203.x

Hewitt JK, Silberg JL, Neale MC, Eaves LJ, Erickson M. The analysis of parental ratings of children's behavior using LISREL. Behavior Genetics. 1992; 22:293–317.10.1007/BF01066663 [PubMed: 1616461]

Hoffman B, Lance CE, Bynum B, Gentry WA. Rater source effects are alive and well after all. Personnel Psychology. 2010; 63:119–151.10.1111/j.1744-6570.2009.01164.x

Holzinger KJ, Swineford F. The bi-factor method. Psychometrika. 1937; 2:41–54.10.1007/BF02287965

Horton NJ, Fitzmaurice GM. Regression analysis of multiple source and multiple informant data from complex survey samples. Statistics in Medicine. 2004; 23:2911–2933.10.1002/sim.1879 [PubMed: 15344194]

Hussong AM, Jones DJ, Stein GL, Baucom DH, Boeding S. An Internalizing Pathway to Alcohol Use and Disorder. Psychology of Addictive Behaviors. 2011; 25:390–404.10.1037/a0024519 [PubMed: 21823762]

Kenny, DA. Correlation and causality. New York: Wiley; 1979.

Kenny DA. A general model of consensus and accuracy in interpersonal perception. Psychological Review. 1991; 98:155–163.10.1037/0033-295X.98.2.155 [PubMed: 2047511]

Kenny DA, Kashy DA. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. Psychological Bulletin. 1992; 112:165–172.10.1037/0033-2909.112.1.165

Kenny, DA.; Kashy, DA.; Cook, WL. Dyadic data analysis. New York: Guilford Press; 2006.

Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ. A new approach to integrating data from multiple informants in psychiatric assessment and research: mixing and matching contexts and perspectives. American Journal of Psychiatry. 2003; 160:1566–1577.10.1176/appi.ajp.160.9.1566 [PubMed: 12944328]

Lance CE, Hoffman BJ, Gentry WA, Baranik LE. Rater source factors represent important subcomponents of the criterion construct space, not rater bias. Human Resource Management Review. 2008; 18:223–232.10.1016/j.hrmr.2008.03.002

Lance CE, Noble CL, Scullen SE. A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. Psychological Methods. 2002; 7:228–244.10.1037/1082-989X.7.2.228 [PubMed: 12090412]

Loken E. Identification constraints and inference in factor models. Structural Equation Modeling. 2005; 12:232–244.10.1207/s15328007sem1202_3

MacCallum RC, Roznowski M, Necowitz LB. Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychological Bulletin. 1992; 111:490–504.10.1037/0033-2909.111.3.490 [PubMed: 16250105]

Marsh HW. Confirmatory factor analyses of multitrait-multimethod data: many problems and a few solutions. Applied Psychological Measurement. 1989; 13:335–361.10.1177/014662168901300402

Marsh HW. Multitrait-multimethod analyses: inferring each trait-method combination with multiple indicators. Applied Measurement in Education. 1993; 6:49–81.10.1207/s15324818ame0601_4

Marsh HW, Bailey M. Confirmatory factor analyses of multitrait-multimethod data: A comparison of the behavior of alternative models. Applied Psychological Measurement. 1991; 15:47–70.10.1177/014662169101500106

Marsh, HW.; Grayson, D. Latent variable models of multitrait-multimethod data. In: Hoyle, RH., editor. Structural Equation Modeling: Concepts, Issues, and Applications. Thousand Oaks, CA: Sage; 1995. p. 177-198.

Marsh HW, Hocevar D. A new, more powerful approach to multitrait-multimethod analyses: application of second-order confirmatory factor analysis. Journal of Applied Psychology. 1988; 73:107–117.10.1037/0021-9010.73.1.107

Mount MK, Judge TA, Scullen SE, Sytsma MR, Hezlett SA. Trait, rater and level effects in 360-degree performance ratings. Personnel Psychology. 1998; 51:557–576.10.1111/j.1744-6570.1998.tb00251.x

Muthén BO. Contributions to factor analysis of dichotomous variables. Psychometrika. 1978; 43:551–560.10.1007/BF02293813

Muthén, BO.; du Toit, SHC.; Spisic, D. Unpublished manuscript. 1997. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.

Muthén, LK.; Muthén, BO. Mplus User's Guide. 6. Los Angeles, CA: Muthén & Muthén; 1998–2010.

Najman JM, Williams GM, Nikels J, Spence S, Bor W, O'Callaghan M, Le Brocque R, Andersen MJ. Mothers' mental illness and child behavior problems: cause-effect association or observation bias? Journal of the American Academy of Child and Adolescent Psychiatry. 2000; 39:592–602.10.1097/00004583-200005000-00013 [PubMed: 10802977]

Neale MC, Stevenson J. Rater bias in the EASI temperament scales: a twin study. Journal of Personality and Social Psychology. 1989; 56:446–455.10.1037/0022-3514.56.3.446 [PubMed: 2926639]

Nussbeck FW, Eid M, Geiser C, Courvoisier DS, Lischetzke T. A CTC(*M*-1) model for different types of raters. Methodology. 2009; 5:88–98.10.1027/1614-2241.5.3.88

Nussbeck FW, Eid M, Lischetzke T. Analysing multitrait-multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: What sample size is needed for valid results? British Journal of Mathematical and Statistical Psychology. 2006; 59:195–213.10.1348/00071100SX67490 [PubMed: 16709286]

Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika. 1979; 44:443–460.10.1007/BF02296207

Pohl S, Steyer R. Modeling common traits and method effects in multitrait-multimethod analysis. Multivariate Behavioral Research. 2010; 45:45–72.10.1080/00273170903504729

Pohl S, Steyer R, Kraus K. Modeling method effects as individual causal effects. Journal of the Royal Statistical Society, Series A. 2008; 171:41–63.10.1111/j.1467-985X.2007.00517.x

Reininghaus U, McCabe R, Burns T, Croudace T, Priebe S. Measuring patients' views: a bifactor model of distinct patient-reported outcomes in psychosis. Psychological Medicine. 2011; 41:277–289.10.1017/S0033291710000784 [PubMed: 20406529]

Reise SP, Ventura J, Keefe RSE, Baade LE, Gold JM, Green MF, Kern RS, Mesholam-Gately R, Nuechterlein KH, Seidman LJ, Bilder R. Bifactor and item response theory analyses of interviewer

report scales of cognitive impairment in schizophrenia. Psychological Assessment. 2011; 23:245–261.10.1037/a0021501 [PubMed: 21381848]

Renk K. Cross-informant ratings of the behavior of children and adolescents: the “gold standard. Journal of Child and Family Studies. 2005; 14:457–468.10.1007/s10826-005-7182-2

Rowe DC, Kandel D. In the eye of the beholder? Parental ratings of externalizing and internalizing symptoms. Journal of Abnormal Child Psychology. 1997; 25:265–275.10.1023/A:1025756201689 [PubMed: 9304443]

Satorra, A. Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In: DDH; Pollock, DSG.; Satorra, A., editors. Innovations in Multivariate Statistical Analysis: A Festschrift for Heinz Neudecker. Heijmans. Kluwer Academic Publishers; Dordrecht: 2000. p. 233-247.

Satorra, A.; Bentler, PM. A scaled difference chi-square test statistic for moment structure analysis. 1999 Aug 3. (Technical Report). Retrieved from UCLA Department of Statitsics website, http://statistics.ucla.edu/preprints/uclastat-preprint-1999:19

Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods. 2002; 7:147–177.10.1037/1082-989X.7.2.147 [PubMed: 12090408]

Scullen SE, Mount MK, Goff M. Understanding the latent structure of job performance ratings. Journal of Applied Psychology. 2000; 85:956–970.10.1037/0021-9010.85.6.956 [PubMed: 11125659]

Seiffge-Krenke I, Kollmar F. Discrepancies between mothers’ and fathers’ perceptions of sons’ and daughters’ problem behaviour: A longitudinal analysis of parent-adolescent agreement on internalising and externalising problem behaviour. Journal of Clinical Psychology and Psychiatry. 1998; 39:687–697.10.1017/S0021963098002492

Simonoff E, Pickles A, Hewitt J, Silberg J, Rutter M, Loeber R, Meyer J, Neale M, Eaves L. Multiple raters of disruptive child behavior: using a genetic strategy to examine shared views and bias. Behavior Genetics. 1995; 25:311–326.10.1007/BF02197280 [PubMed: 7575360]

Skrondal, A.; Rabe-Hesketh, S. Generalized latent variable modeling: Multilevel, longitudinal and structural equation models. Boca Raton, FL: Chapman & Hall/CRC; 2004.

Spearman C. General intelligence, objectively determined and measured. American Journal of Psychology. 1904; 15:201–293.10.2307/1412107

Steyer R. Models of classical psychometric test theory as stochastic measurement models: representation, uniqueness, meaningfulness, identifiability and testability. Methodika. 1989; 3:25–60.

Steyer R, Ferring D, Schmitt MJ. States and traits in psychological assessment. European Journal of Psychological Assessment. 1992; 8:79–98.

Thissen, D.; Orlando, M. Item response theory for items scored in two categories. In: Thissen, D.; Wainer, H., editors. Test scoring. Mahwah, NJ: Erlbaum; 2001. p. 73-140.

Thomson GH. Some points of mathematical technique in the factorial analysis of ability. Journal of Educational Psychology. 1936; 27:37–54.10.1037/h0062007

Thomson, GH. The factorial analysis of human ability. 5. Boston: Houghton Mifflin; 1951.

Thurstone, LL. Vectors of the mind. Chicago, IL: University of Chicago Press; 1935.

van der Valk JC, van den Oord EJCG, Verhulst FC, Boomsma DI. Using parental ratings to study the etiology of 3-year-old twins’ problem behaviors: different views or rater bias? Journal of Child and Adolescent Psychiatry. 2001; 42:921–931.10.1111/1469-7610.00788

van Dulmen MHM, Egeland B. Analyzing multiple informant data on child and adolescent behavior problems: predictive validity and comparison of aggregation procedures. International Journal of Behavioral Development. 2011; 35:84–92.10.1177/0165025410392112

Widaman KF. Hierarchically nested covariance structure models for multitrait-multimethod data. Applied Psychological Measurement. 1985; 9:1–26.10.1177/014662168500900101

Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. Psychological Methods. 2007; 12:58–79.10.1037/1082-989X.12.1.58 [PubMed: 17402812]

Woehr DJ, Sheehan MK, Bennett W Jr. Assessing measurement equivalence across rating sources: A multitrait-multirater approach. Journal of Applied Psychology. 2005; 90:592–600.10.1037/0021-9010.90.3.592 [PubMed: 15910153]

Yoon M, Millsap RE. Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. Structural Equation Modeling. 2007; 14:435–463.10.1080/10705510701301677

Youngstrom EA, Izard C, Ackerman B. Dysphoria-related bias in maternal ratings of children. Journal of Consulting and Clinical Psychology. 1999; 68:1038–1050.10.1037/0022-006X.67.6.905 [PubMed: 11142538]

Zucker RA, Ellis DA, Bingham CR, Fitzgerald HE, Sanford KP. Other evidence for at least two alcoholisms, II: Life course variation in antisociality and heterogeneity of alcoholic outcome. Development and Psychopathology. 1996; 8:831–848.10.1017/S0954579400007458

Zucker, RA.; Fitzgerald, HE.; Refior, SK.; Puttler, LI.; Pallas, DM.; Ellis, DA. The clinical and social ecology of childhood for children of alcoholics: Description of a study and implications for a differentiated social policy. In: Fitzgerald, HE.; Lester, BM.; Zuckerman, BS., editors. Children of addiction: Research, health, and policy issues. New York, NY: Routledge Falmer; 2000. p. 109-141.

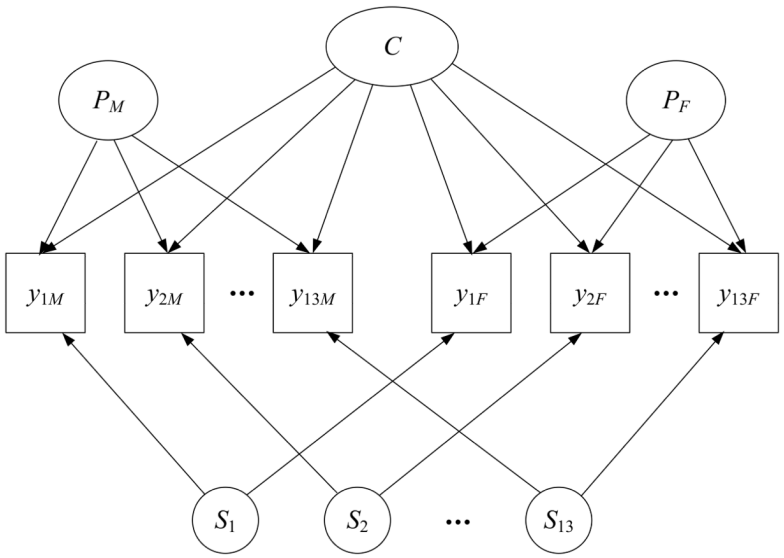

Figure 1.
Unconditional tri-factor model for parent-report ratings on 13 items. The observed ratings are numbered by item; *M* and *F* subscripts differentiate ratings of the mother and father, respectively. Intercepts and random error terms are not shown in the diagram.

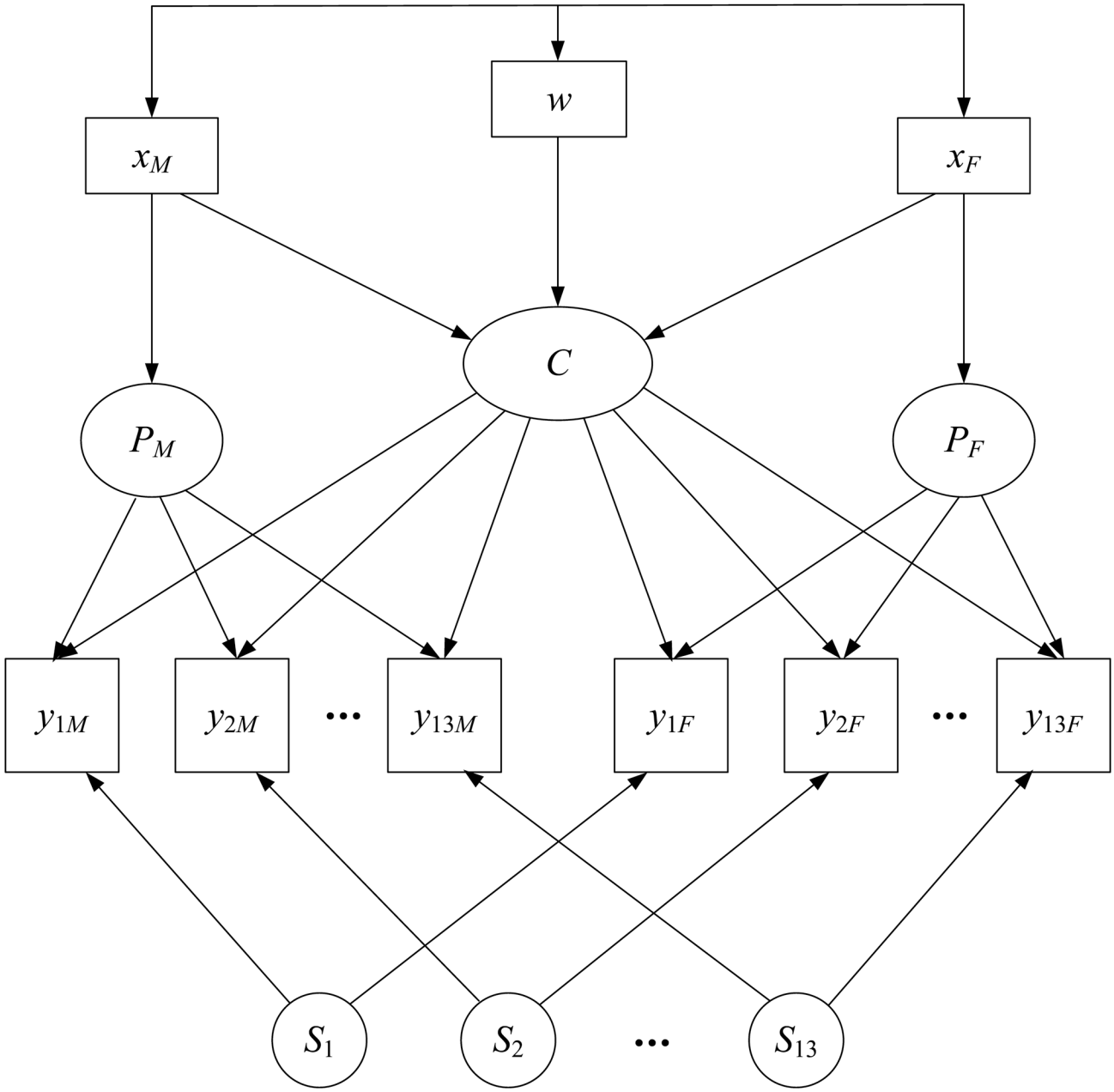

Figure 2.
Conditional tri-factor model for parent-report ratings on 13 items. The observed ratings are numbered by item; *M* and *F* subscripts differentiate ratings of the mother and father, respectively. The predictor *w* is a target characteristic and the predictors x_M and x_F are informant-specific predictors. Intercepts and error terms/disturbances are not shown in the diagram.

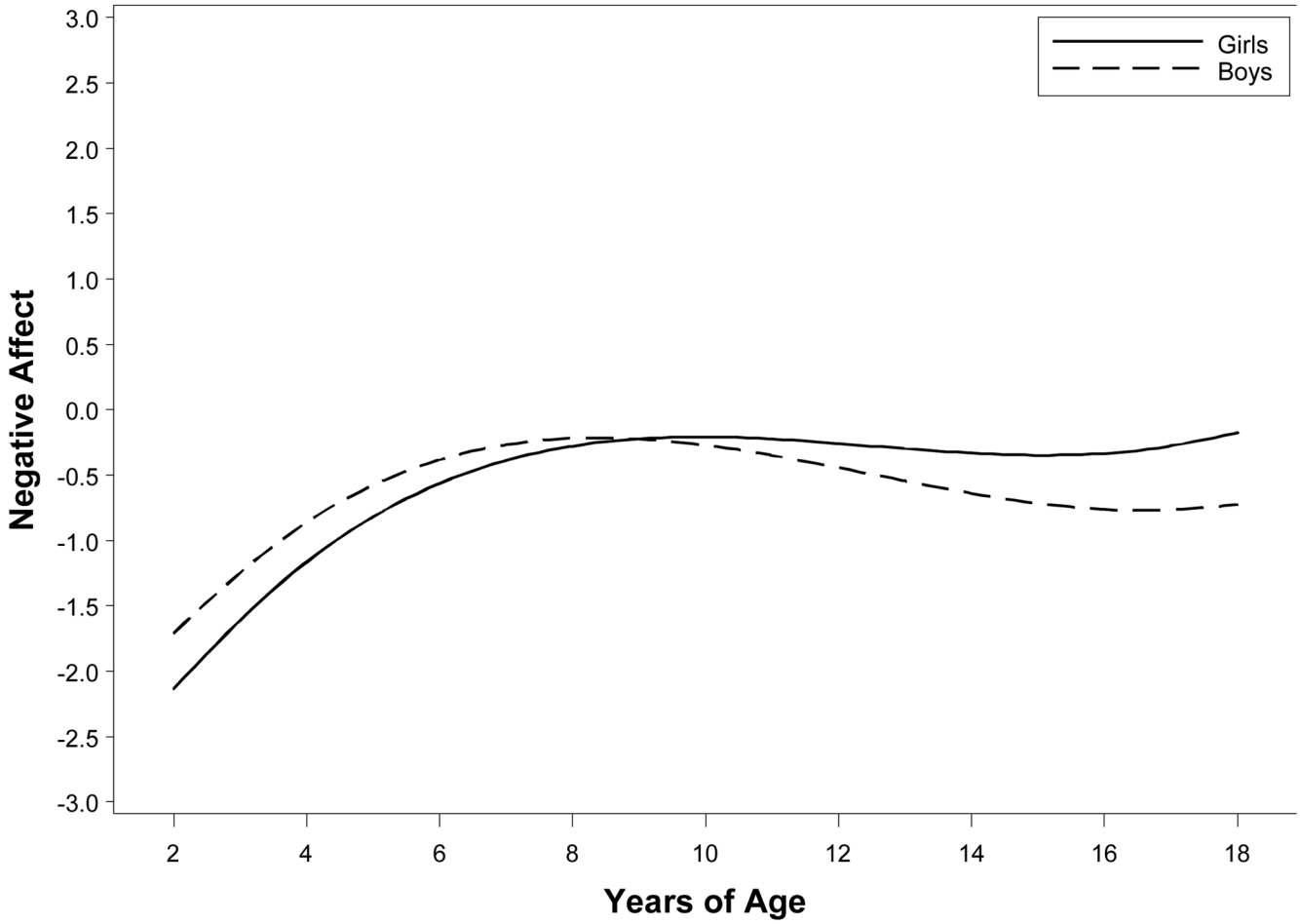

Figure 3.
Age trends in the common factor for negative affect for girls and boys (averaging over other predictors).

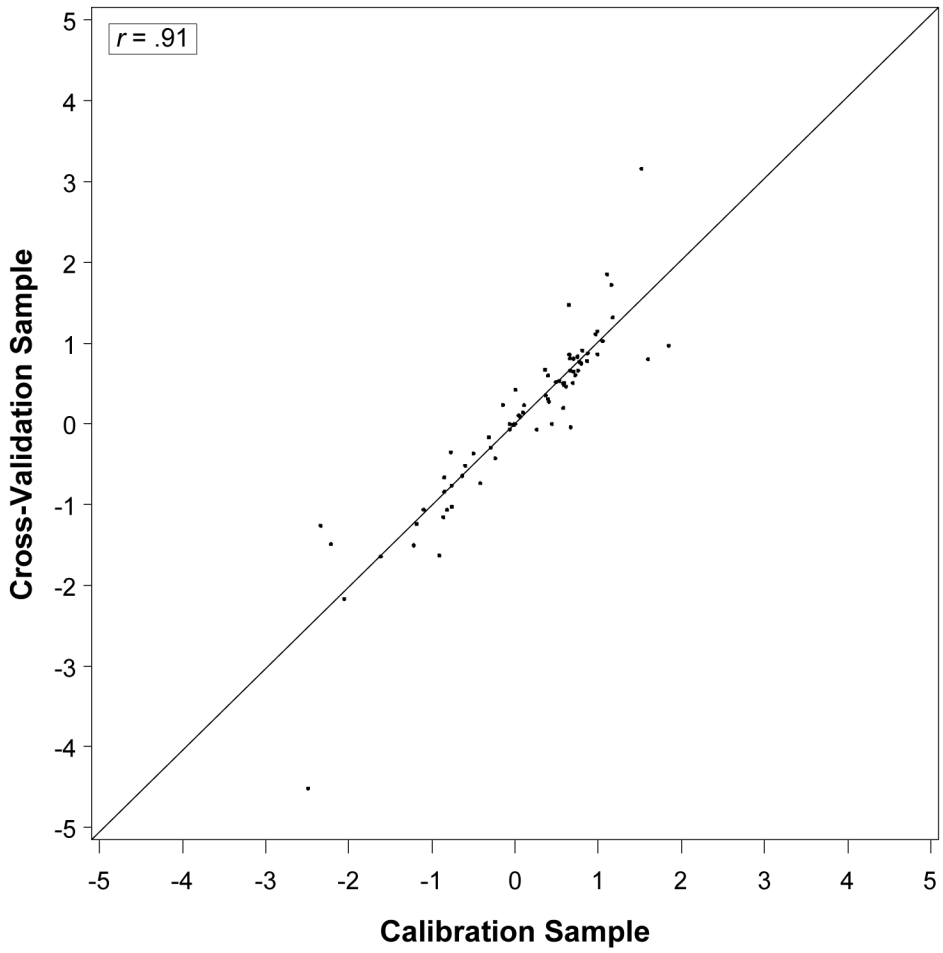

Figure 4.
Plot illustrating stability of parameter estimates across two cross-sectional draws from the pooled longitudinal data.

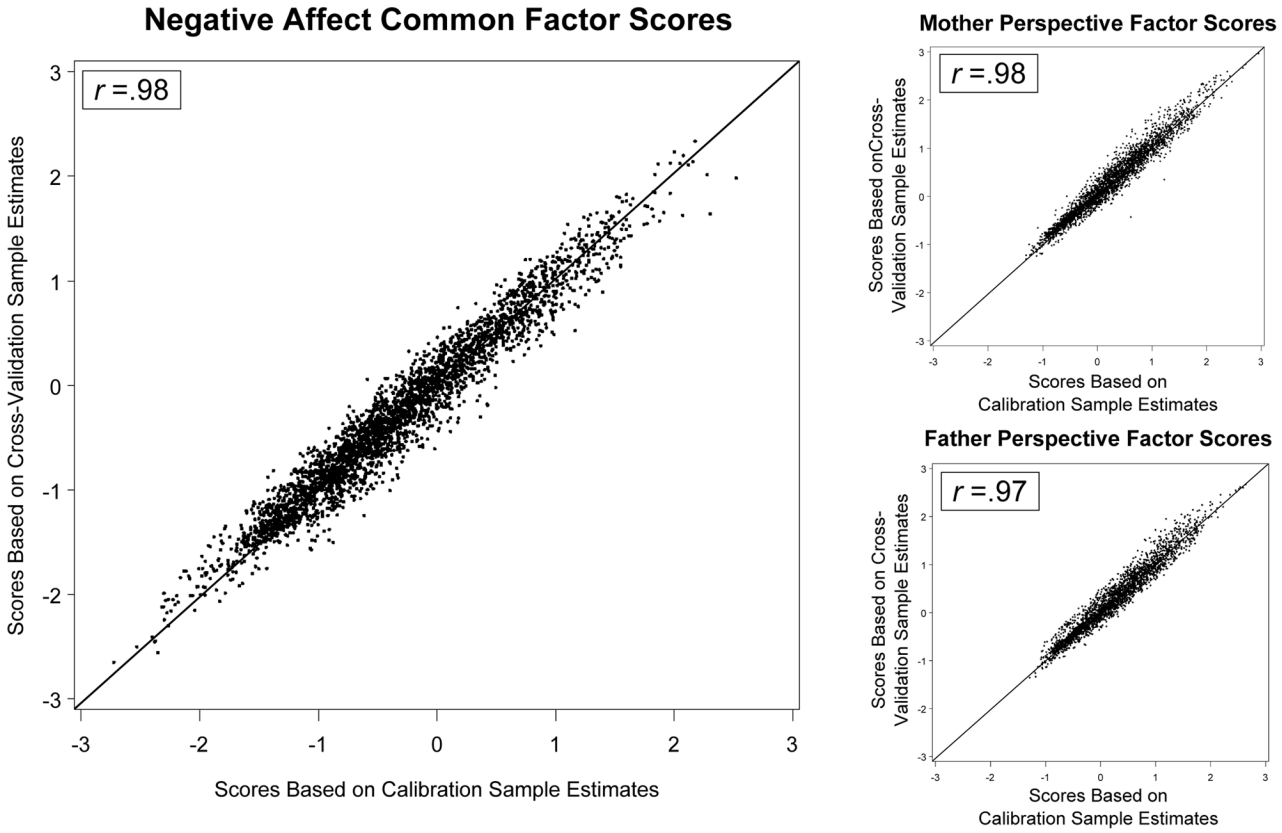

Figure 5.
Scatterplots illustrating the stability of MAP scores generated from estimates obtained from two non-overlapping cross-sectional draws from the longitudinal data.

Table 1

Number of observations by age and study

	Full Longitudinal Sample		Calibration Sample		Cross-Validation Sample	
Age	**MLS**	**AFDP**	**MLS**	**AFDP**	**MLS**	**AFDP**
2	11	-	2	-	2	-
3	137	-	45	-	24	-
4	111	-	27	-	27	-
5	91	-	30	-	20	-
6	127	-	29	-	29	-
7	122	-	31	-	29	-
8	107	-	25	-	30	-
9	160	-	41	-	42	-
10	150	32	43	9	38	12
11	131	107	36	42	31	27
12	172	191	55	55	55	73
13	154	266	45	96	48	82
14	149	294	48	89	43	103
15	159	247	54	84	33	87
16	146	150	54	64	32	42
17	139	54	50	15	35	19
18	25	4	9	0	8	4
N	2091	1345	626	454	526	449

Table 2

Numbers of non-missing values (N) and endorsement rates for negative affect items in the calibration sample.

	Mother		Father	
Item	**N**	**endorse**	**N**	**endorse**
1. lonely	1048	0.17	905	0.18
2. cries a lot	1048	0.17	906	0.14
3. fears do something bad	1047	0.16	904	0.20
4. has to be perfect	1049	0.41	905	0.39
5. no one loves him/her	1048	0.23	906	0.18
6. worthless or inferior	1049	0.22	904	0.16
7. nervous/tense	1049	0.21	906	0.22
8. fearful/anxious	1048	0.18	906	0.16
9. feels guilty	1049	0.11	906	0.09
10. sulks a lot	1047	0.27	907	0.25
11. sad/depressed	1048	0.23	907	0.19
12. worries	1048	0.38	905	0.29
13. others out to get him/her	1048	0.10	906	0.10

Table 3

Raw and standardized intercept and factor loading estimates from the conditional tri-factor model fit to the calibration sample (final model).

Raw Estimates

		Factor Loading		
Item	**Intercept**	**Common**	**Perspective**	**Specificity**
1. lonely	−0.85	0.73	0.62	0.49
2. cries a lot	−0.60	0.65	0.59	0.81
3. fears do something bad	−0.82 (M), −0.76 (F)	0.41	0.70	0.59
4. has to be perfect	0.45	0.11	0.78	0.80
5. no one loves him/her	−0.42 (M), −1.08 (F)	0.97	0.71	0.40
6. worthless or inferior	−1.18 (M), −1.61 (F)	1.06	1.00	0.67
7. nervous/tense	−2.34	0.66	1.60	1.85
8. fearful/anxious	−1.18	0.37	1.16	0.65
9. feels guilty	−2.49	0.87	1.52	1.11
10. sulks a lot	−0.50	0.66	0.70	0.00
11. sad/depressed	−1.22	0.99	0.87	0.01
12. worries	0.26 (M), −0.24 (F)	0.58	1.18	0.76
13. others out to get him/her	−2.05	0.76	0.67	0.40

Standardized Solution

		Factor Loading		
Item	**Intercept**	**Common**	**Perspective**	**Specificity**
1. lonely	−0.53	0.52	0.39	0.50
2. cries a lot	−0.33	0.41	0.33	0.69
3. fears do something bad	−0.55 (M), −0.51 (F)	0.31	0.48	0.43
4. has to be perfect	0.28	0.08	0.50	0.58
5. no one loves him/her	−0.24 (M), −0.64 (F)	0.65	0.42	0.23
6. worthless or inferior	−0.59 (M), −0.81 (F)	0.60	0.51	0.34
7. nervous/tense	−0.84	0.27	0.59	0.67
8. fearful/anxious	−0.62	0.22	0.62	0.50
9. feels guilty	−1.00	0.40	0.62	0.49
10. sulks a lot	−0.35	0.52	0.49	0.00
11. sad/depressed	−0.69	0.64	0.51	0.00
12. worries	0.14 (M), −0.12 (F)	0.35	0.63	0.43
13. others out to get him/her	−1.33	0.56	0.44	0.26

Note: Intercepts that differed between mothers and fathers are labeled (M) or (F), respectively, to specify the informant. Standardized factor loading estimates differ slightly between mothers and fathers (never exceeding a difference of .04). Values for mothers are reported.

Table 4

Effects obtained from the regression of common, perspective, and specific factors on age, gender (male), study (MLS) and parental impairment.

Effect	Estimate	SE	Partially Standardized Estimate
Common Factor			
Age	−0.003	0.042	−0.003
Age^2	**−0.015**	0.003	**−0.013**
Age^3	**0.002**	0.001	**0.002**
MLS	**−0.764**	0.131	**−0.669**
Male	−0.063	0.128	−0.055
Male*Age	**−0.061**	0.031	**−0.054**
Parent alcoholism	0.095	0.088	0.083
Parent alcoholism + ASP	**0.582**	0.174	**0.510**
Parent depression	**0.369**	0.130	**0.323**
Perspective Factors			
Parent alcoholism	0.061	0.096	0.060
Parent alcoholism + ASP	−0.145	0.179	−0.143
Parent depression	**0.538**	0.126	**0.528**
Specific Factors			
1. Lonely			
Age	**−0.313**	0.119	**−0.196**
2. Cries a lot			
Age	**−0.295**	0.079	**−0.191**
Male	**−0.633**	0.197	**−0.411**
3. Fears do something bad			
MLS	**−0.860**	0.289	**−0.796**
4. Has to be perfect			
Age	0.047	0.049	0.041
Age^2	**−0.019**	0.005	**−0.017**
MLS	**−0.771**	0.191	**−0.672**
8. Fearful/anxious			
MLS	**−2.210**	0.597	**−1.522**
9. Feels guilty			
MLS	**−0.912**	0.285	**−0.836**
12. Worries			
MLS	**−0.847**	0.239	**−0.786**

Note. Bold entries are significant at $p < .05$. Partially standardized estimates are computed by standardizing the factors but leaving regressors in their raw scales. Age was centered at 10 years. All binary predictors are named to indicate the presence of the characteristic (e.g., MLS is scored 1 for targets from the MLS study, 0 for targets from the AFDP study).

Table 5

Correlations between tri-factor analysis common factor scores, the average proportion of items endorsed by mothers and fathers, the proportion of items endorsed by either mothers or fathers, and the average of factor score estimates obtained separately for mothers and fathers.

	TriFS	AvgProp	OrProp	AvgFS1	AvgFS2
Tri-factor common factor scores (TriFS)	1.00				
Average proportion of items endorsed by mothers and fathers (AvgProp)	0.79	1.00			
Proportion of items endorsed by either mothers or fathers (OrProp)	0.79	0.95	1.00		
Average of factor scores estimates obtained separately for mothers and fathers from a 2PL-IRT model (AvgFS1)	0.82	0.97	0.93	1.00	
Average of factor scores estimates obtained separately for mothers and fathers from a MNLFA model (AvgFS2)	0.84	0.97	0.93	0.99	1.00

Note. 2PL-IRT = two-parameter logistic item response theory model, MNLFA model = moderated nonlinear factor analysis model