CHAPTER FIVE

The Incorporation of Categorical Measurement Models in the Analysis of Individual Growth

Patrick J. Curran
Michael C. Edwards
R. J. Wirth
Andrea M. Hussong
University of North Carolina at Chapel Hill

Laurie Chassin
Arizona State University

The empirical study of human development is a daunting task, particularly when focusing on the first few decades of life when change is both rapid and variable. The fields of child development and developmental psychopathology are supported by rich and dynamic theoretical models that strive to capture the complex and subtle processes of individual growth over time. For example, developmental theory is built upon core tenets such as homotypic continuity (Kagan, 1971), developmental coactions and transactions (e.g., Gottlieb & Halpern, 2002; Sameroff, 1995), and dynamic processes of multifinality and equifinality (e.g., Cicchetti & Rogosch, 1996; Gottlieb, Wahlsten, & Lickliter, 1998), to name a few. However, it is often challenging, if not impossible, to fully test these complex theories from an empirical perspective (e.g., Curran & Willoughby, 2003; Wohlwill, 1970, 1973). A core issue is the need to select a statistical model that optimally corresponds to the theoretical model that gave rise to the proposed research hypotheses.

The degree to which the theoretical model diverges from the statistical model directly impacts the validity of our empirically based conclusions (Curran & Willoughby, 2003; Curran & Wirth, 2004; Wohlwill, 1970). We must take great care in selecting a statistical model that optimally corresponds to the theoret-

ical model under study. This issue is particularly salient when considering developmental processes both within and across different contexts. Although there are a large number of components that must be considered when selecting an appropriate statistical model to test a given theory of interest, we focus on three specific issues here: The need to explicitly incorporate repeated measures that are discretely scaled (e.g., dichotomous or ordinal); to differentially weight individual items when forming a scale as a function of item reliability and severity; and to incorporate optimally-derived scale scores in growth curve models of developmental stability and change over time.

Our motivating goal is to describe and empirically demonstrate a two-stage procedure of scale construction and growth curve analysis that addresses challenges encountered when studying individual and contextual influences on development. The first stage involves the calculation of individual and time specific scale scores based on a set of dichotomous repeated measures. We consider three scoring methods: proportion scores, item response theory, and categorical confirmatory factor analysis. The second stage involves incorporating these scale scores into a general multilevel modeling framework for the analysis of developmental trajectories over time. Although we focus our attention on the multilevel (or hierarchical linear) model, all of our arguments extend directly to the structural equation based latent curve model as well (e.g., Bollen & Curran, 2006). Our expectation is that the development and application of a proper measurement model in the first stage will improve the validity and reliability of the growth curve models fitted in the second stage.

We begin with a brief review of the standard linear multilevel growth curve model and highlight the assumptions of continuous outcomes and the modeling of a single score over time. We then summarize existing methods for the simultaneous estimation of a growth curve model with a multiple item measurement model. Next we present three existing measurement models for categorical repeated measures data: the proportion score, the 2-parameter logistic (2PL) item response theory model, and the categorical confirmatory factor analysis model. We then demonstrate the comparative utility of these models by examining multiple repeated assessments of externalizing symptomatology in a sample of 444 adolescents ranging in age from 10 to 21. We conclude with potential expansions of these models and recommendations for the use of these techniques in practice.

MULTILEVEL GROWTH CURVE MODELS

The estimation of longitudinal growth curves has been an interest in the social sciences for nearly two centuries (see Bollen & Curran, 2006, for a historical

review). There are several powerful methodological frameworks that can be applied to the estimation of growth curves based on repeated measures data. Two widely used approaches are the structural equation modeling based latent curve analysis (LCA; Bollen & Curran, 2006; McArdle, 1988, 1989; Meredith & Tisak, 1984, 1990), and the multilevel (or mixed) modeling framework (e.g., Bryk & Raudenbush, 1987; Raudenbush & Bryk, 2002; Willett & Sayer, 1994). There are many important elements shared between these two modeling approaches, but several key differences remain (Bauer, 2003; Curran, 2003; Raudenbush, 2001; Willett & Sayer, 1994).

Multilevel modeling is a general analytic framework that allows for the explicit modeling of nested (or nonindependent) data structures (e.g., Goldstein, 1986; Mason, Wong, & Entwisle, 1983; Raudenbush & Bryk, 2002). Classic examples include children nested within families, classrooms nested within schools, and households nested within neighborhoods. However, repeated assessments over time can be conceptualized as nested within individuals, and thus the multilevel model can be directly applied to the analysis of growth curves (Bryk & Raudenbush, 1987).

For the standard linear multilevel growth curve model, we can define a level-1 or within-person equation, and a level-2 or between-person equation. We define the level-1 equation as:

$$y_{ti} = \beta_{0i} + \beta_{1i}x_{ti} + e_{ti} \ , \tag{1}$$

where the outcome variable y assessed at time t for individual i can be expressed as an additive function of an intercept (β_{0i}), a linear slope (β_{1i}) multiplied by the value of time at assessment t for individual i (x_{ti}), and a time- and individual-specific residual (e_{ti}). Because the individually varying intercepts and slopes are treated as random variables, these can be expressed as:

$$\beta_{0i} = \gamma_{00} + u_{0i} \tag{2}$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \ , \tag{3}$$

where γ_{00} and γ_{10} represent the mean intercept and slope, respectively, and u_{0i} and u_{1i} represent individual deviations from these means.

The level-1 and level-2 distinction is for heuristic value only. Equations 2 and 3 can thus be substituted into Equation 1 to result in the reduced form expression

$$y_{ti} = (\gamma_{00} + \gamma_{10}x_{ti}) + (u_{0i} + u_{1i}x_{ti} + e_{ti}) \ . \tag{4}$$

The parameters of interest are the fixed and random effects associated with Equation 4. Specifically, the fixed effects (γ_{00} and γ_{10}) represent the mean intercept and mean slope pooling over all individuals. The random effects

include the variance of the residuals at level-1 (denoted $var\,(e_{ti}) = \sigma^2$), and the variance of the individual deviations around the mean intercept and slope (denoted $var\,(u_{0i}) = \tau_{00}$ and $var\,(u_{1i}) = \tau_{11}$, respectively). It is common to assume that the level-1 residual variance is homoscedastic and independent over time (e.g., $\sigma^2\mathbf{I}$), although this restriction can be tested and relaxed if needed. It is also common to estimate the covariance between the intercepts and slopes (denoted $cov\,(u_{0i}, u_{1i}) = \tau_{01}$), which can in turn be rescaled as a correlation coefficient (but see Biesanz, Deeb-Sossa, Aubrecht, Bollen, & Curran, 2004, for details of the scale-dependence of this covariance).

There are a variety of interesting ways in which this model can be expanded. For example, Equation 1 can include powered terms of time (e.g., x_{ti}^2) to estimate more complex members of the polynomial trajectory family (e.g., quadratic, cubic). Similarly, two measures of time could be defined that allow for the estimation of linear splines connected at a knot point (Raudenbush & Bryk, 2002). Time varying covariates (TVCs) can be incorporated in Equation 1 to include time-specific predictors of the repeated measures net the influence of the underlying random trajectories (e.g., days of school missed per year, onset of a new medical diagnosis, recent alcohol consumption). Time invariant covariates (TICs) can be incorporated in Equations 2 and 3 to predict individual variability in intercepts and slopes (e.g., gender, ethnicity, country of origin). The multilevel model is characterized by a multitude of significant strengths and has been widely used in the analysis of repeated measures data.

Despite these strengths, there are several issues of which we must be aware when considering the application of these techniques to developmental data. First, the standard multilevel model described earlier assumes that the residuals are continuously and normally distributed; however, many outcomes in developmental research are dichotomous, ordinal, or count variables which directly violate this distributional assumption. Second, the standard two-level model assumes that the outcome is a single individual and time specific score (i.e., y_{ti}). However, as developmental researchers, we are often not interested in a single score, but instead would like to incorporate a psychometric scale consisting of multiple items (e.g., scales assessing internalizing symptomatology or delinquent behavior), and to then fit growth models to these measurement models. The inclusion of multiple item psychometric models not only increases the content validity of the assessment of our construct, but also increases reliability and power (e.g., Bollen, 1989).

The purpose of our chapter is to present alternative methods that can be used to create individual and time specific scale scores from a set of dichotomous measures for use in multilevel models. From a statistical standpoint, the ideal approach would estimate the measurement model and the growth model simultaneously. However, as we discuss in greater detail next, these single-stage

methods are often not easily applied within developmental research settings. Prior to presenting our two-stage analytic strategy, we briefly review existing single-stage approaches to this problem.

Single-Stage Categorical Growth Curve Models

As noted earlier, the standard multilevel model typically assumes that the repeated measures are continuously (and often normally) distributed. However, for many areas within developmental research, the repeated measures of interest are often discretely scaled (e.g., dichotomous or polytomous). This is particularly evident in studies of developmental psychopathology in which child outcomes are measured in terms of specific symptoms, discrete behaviors, or diagnostic status. It is well established that the application of our standard measurement and growth curve models to measures that are discretely scaled introduces potentially significant bias in the analysis and subsequent inferences (Mehta, Neale, & Flay, 2004; Muthén & Kaplan, 1985). Therefore, we must carefully consider the scaling of the outcome measure when selecting the optimal analytic method to test our research hypotheses. Fortunately, there are well developed modeling strategies that allow for the explicit incorporation of categorical repeated measures data in both measurement and growth curve models. There are two important approaches currently available.

The first method for the simultaneous estimation of categorical repeated measures in growth curve analysis is nonlinear multilevel models (e.g., Davidian & Giltinan, 1995; Diggle, Heagerty, Liang, & Zeger, 2002; Gibbons & Hedeker, 1997; Zeger, Liang, & Albert, 1988). Whereas the linear multilevel model imposes an identity link function for the analysis of continuous dependent measures, the nonlinear multilevel model incorporates a variety of alternative link functions for the analysis of dichotomous, ordinal, and count data (e.g., McCullagh & Nelder, 1989). Although well developed for single item outcomes, these models become quite complex when fitting measurement models to a set of items hypothesized to define an underlying construct (e.g., multiple dichotomous items assessing childhood aggression). Furthermore, it is often quite difficult to achieve numerical convergence of nonlinear multilevel models when fitted to empirical data of the type commonly encountered in developmental research (e.g., small sample sizes, scales consisting of multiple dichotomous items, many repeated assessments, attrition over time). An important related alternative is the three-level nonlinear multilevel Rasch model proposed by Raudenbush, Johnson, and Sampson (2003). This method is both creative and powerful, but also imposes certain restrictions to achieve identification, restrictions that might not hold in many areas of developmental research. In sum, these nonlinear approaches are both analytically elegant and highly promising but are

currently infeasible for testing many developmental questions of interest.

The second available method for fitting growth curve models is the nonlinear structural equation model (SEM; e.g., Jöreskog, 1994; Muthén, 1983; 1984). This approach is conceptually similar to the nonlinear multilevel model, although the estimation procedures differ in important ways (Wirth & Edwards, 2005). As we describe in greater detail later, the nonlinear SEM is based on the premise that the observed dichotomous measures are discrete realizations of a truly continuous unobserved response distribution. The goal is to simultaneously estimate the correlation structure among the underlying distributions and to fit the growth curve model of interest (e.g., Mehta et al., 2004). As with the nonlinear multilevel model, the nonlinear SEM is both powerful and flexible. However, this approach can be characterized by significant convergence and estimation problems when fitted to empirical data of the type commonly encountered in developmental research. Finally, as with the three-level Rasch model of Raudenbush et al. (2003), restrictive constraints are similarly needed to identify a growth model fitted to dichotomous scales (Muthén, 1996). Taken together, although the nonlinear SEM is a highly promising analytic strategy, this too is currently limited when simultaneously fitting measurement and growth curve models to developmental data.

In sum, there are a number of existing methodologies for the simultaneous analysis of psychometric measurement models fitted to dichotomous repeated measures combined with a random coefficient growth curve model. We have briefly reviewed the nonlinear multilevel model and the nonlinear SEM, although there are several other approaches we do not detail here (e.g., the random coefficient 2PL item response model; Fox, 2003, 2005). Despite both great flexibility and promise, these simultaneous estimation techniques are often empirically intractable given the many complexities encountered in developmental research settings. For this reason, we instead approach the problem from a two-stage perspective in which we separate the fitting of the psychometric model from the fitting of the growth curve model. This two-stage approach is not ideal from a statistical efficiency perspective given that not all model parameters are estimated simultaneously and there is thus some loss of statistical information. However, we view this strategy as a pragmatic alternative to the more elegant, yet analytically less tractable, single-stage methods.

THREE CATEGORICAL MEASUREMENT MODELS

Proportion Scoring

Arguably the most widely used method for imposing a measurement model on a set of categorical (and typically dichotomous) repeated measures is the proportion score.[1] Here, the manifest (or observed) individual and time specific variable p_{ti} is calculated as the proportion of items endorsed positively relative to the total number of available items. More formally,

$$p_{ti} = \frac{\sum_{j=1}^{J_{ti}} y_{jti}}{J_{ti}} \tag{5}$$

where y_{jti} is the observed score on dichotomous item $j = 1, 2, ..., J_{ti}$ at time $t = 1, 2, ...T$ for individual $i = 1, 2, ..., N$ with possible outcome 0 or 1. The values p_{ti} are computed from Equation 5 and the outcome is the unit of analysis for subsequent modeling.[2] For example, each individual subject i might have endorsed $J_t = 10$ items at time t indicating the presence or absence of ten specific antisocial acts having occurred in the prior 30 days. The individual and time specific measure p_{ti} would thus range from 0 to 1 by increments of .1 and reflect the proportion of items endorsed at each time period. Depending upon the age range under study, we might expect these proportion scores to systematically increase or decrease as a function of time (e.g., Moffitt, 1993).

Potential advantages of this approach include intuitive appeal, ease of implementation, and direct interpretation of the associated metric (e.g., .4 unambiguously reflects that 4 of 10 items were endorsed). However, there are two significant limitations that may well outweigh the potential advantages. First, there is no accounting for variation in the *severity* of items within individual or across time. This is because the endorsement of any given item is weighted equally across all available items (e.g., *lying to adults* and *using a weapon in a fight* are equally weighted in the computation of p_{ti}). Second, this approach assumes that the equal weighting of the set of $J-$items validly and reliably captures the underlying construct across time and development (e.g., *pinching and biting* is assumed to be equally indicative of antisocial behavior for all ages between 5 and 15). It is important to realize that the proportion score is a psychometric measurement model, but one that imposes a number of strict conditions. Violation of these conditions may introduce significant bias in the resulting inferences made from models fitted to measures scored in this way.

[1] Note that the proportion score is simply the mean of a set of dichotomous items. If the number of items is constant over time, the proportion score is a simple linear transformation of the sum of a set of dichotomous items. Thus all of our developments here apply to both proportion scores and sum scores when the item set is constant over time.

[2] We use the triple subscripts of j, t, and i to provide a maximally general framework that allows for the possibility of variations in the number of items over both time and individual.

Item Response Theory

The second categorical measurement model we consider is the item response theory (IRT) model. IRT is a collection of statistical models that formally link individuals and discretely scaled test items. In general, IRT models attempt to explain an observed item response in terms of item parameters and the unobserved examinee level on the underlying trait being measured. Importantly, IRT models are intrinsically nonlinear and thus specifically designed for categorical data.

The two parameter logistic model (2PL) is one of the most widely used IRT models and is conventionally written as

$$P(y_{jti} = 1|\theta_{ti}) = \frac{1}{1 + e^{-1.7a_{jt}(\theta_{ti} - b_{jt})}} , \quad (6)$$

where P indicates probability, y_{jti} is the observed response for item j at time t for individual i, θ_{ti} is the latent construct hypothesized to underlie the observed item response patterns, a_{jt} is the discrimination (or slope) parameter and b_{jt} is the threshold (or severity) parameter for item j at time t, and the constant value 1.7 scales the logistic approximation to the normal ogive model.

Discrimination (i.e., a_{jt}) reflects the degree to which responses to the item distinguish between different levels of the latent variable. Alternately, discrimination can be considered a measure of the degree of relation between a particular item and the underlying construct being measured (with higher values indicating a stronger relationship). *Severity* (i.e., b_{jt}), which is also considered *difficulty* in the context of educational testing, is the location on the underlying dimension at which an individual has a 50% chance of endorsing item j at time t. Readers familiar with IRT will note that the standard 2PL model is typically subscripted for item and not time. However, the inclusion of a subscript for time allows for the potential inclusion of longitudinally varying item parameters.

An alternative to the 2PL IRT model for dichotomous data is the 1-parameter logistic (1PL) model, often referred to as the Rasch model (Bond & Fox, 2001; Rasch, 1960). The equation for this model is identical to Equation 6, with the important exception that the slope parameter does not vary over items (i.e., a_{jt} is constant over all items j and times t). This model implies that all items are equally related to the latent construct. In other words, all items are restricted to be equally discriminating. When this model holds the parameter estimation and interpretation of the resulting parameters are greatly simplified. However, in practice, scales not specifically built to Rasch specifications rarely display Rasch-like properties. For this and other reasons, we focus on the 2PL model for the remainder of the chapter.

Our primary goal here is to use the 2PL IRT model as a first stage analysis to obtain individual and time specific estimates of the underlying latent score, and then to take these scores to our second stage analysis consisting of the multilevel growth model. To accomplish this, we first fitted IRT models for item calibration; this step provides the item-specific a and b parameters linking each item to the underlying latent distributions. We then used these calibration parameters to compute individual- and time-specific scores (denoted $\hat{\theta}$). The calibration stage provides the item parameters necessary for creating scores. No currently implemented estimation procedure allows for the simultaneous estimation of all item and person parameters.

Several methods for scoring are available, and Thissen and Wainer (2001) provide an excellent discussion of the technical details of these calculations. Briefly, each unique response pattern (e.g., each pattern of endorsed and nonendorsed items) is characterized by a posterior distribution which is the product of the population distribution and the trace lines corresponding to the pattern of endorsed/not endorsed items. While an entire posterior is available for each response pattern, it is often more convenient to have point estimates for each individual along with a corresponding measure of precision. Two popular estimates of $\hat{\theta}$ are the *modal a posteriori* (MAP) and the *expected a posteriori* (EAP). For the empirical results we present here, we estimate $\hat{\theta}$ via the MAP procedure.[3]

Categorical Confirmatory Factor Analysis

The third and final measurement model we consider is the categorical confirmatory factor analysis (CCFA) model. Key strengths of the CCFA include the ability to differentially weight items as a function of item discrimination, the explicit modeling of measurement error within and across time, and the provision of formal tests of measurement invariance. There is a close correspondence between the CCFA and the 2PL IRT approaches (e.g., Takane & de Leeuw, 1987), although we do not detail this here. We begin with a brief review of the standard linear confirmatory factor analysis (CFA) model fitted to continuously distributed indicators, and then extend this to the more complicated CCFA model.

The data model for the standard linear CFA is defined as:

$$y_{jti} = v_{y_{jt}} + \lambda_{y_{jt}}\eta_{ti} + \varepsilon_{jti} , \quad (7)$$

[3]It is also possible to compute standard errors for the $\hat{\theta}$ scores, although the formal inclusion of these in the second stage analysis significantly complicates estimation and inference. Future research is needed regarding how to use these standard errors in two-stage analysis.

where y_{jti} is the observed continuously distributed score on item j at time t for individual i. This observed score is expressed as a linear combination of an item and time specific intercept ($v_{y_{jt}}$), a time and individual specific latent score on the hypothesized theoretical construct (η_{ti}), an item and time specific factor loading linking the observed item to the underlying latent factor ($\lambda_{y_{jt}}$), and an item, time and individual specific error (ε_{jti}). Importantly, a variety of equality constraints can be imposed over item and across time to simplify these expressions considerably, and we make use of these constraints later.

Equation 7 defines the data model for our observed repeated measures. The covariance and mean structure implied by Equation 7 as a function of the parameters contained in vector $\boldsymbol{\theta}$ is

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_y \boldsymbol{\Psi}_{\eta\eta} \boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_{\varepsilon\varepsilon} \ , \tag{8}$$

with corresponding mean structure

$$\boldsymbol{\eta} = \boldsymbol{\mu}_{\boldsymbol{\eta}} + \boldsymbol{\zeta}. \tag{9}$$

Here, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ represents the the model-implied covariance matrix for N individuals measured on p variables y. $\boldsymbol{\Lambda}_y$ denotes a $p\times$ m matrix of factor loadings (i.e., regression coefficients), $\boldsymbol{\Psi}_{\eta\eta}$ denotes a $m \times m$ variance/covariance or correlation matrix of latent factors, and $\boldsymbol{\Theta}_{\varepsilon\varepsilon}$ denotes a $p \times p$ matrix of unique variances (and potentially covariances). The m-vector of latent factors, $\boldsymbol{\eta}$, is equal to a m-vector of latent means, $\boldsymbol{\mu}_{\boldsymbol{\eta}}$, plus a m-vector of latent deviations, $\boldsymbol{\zeta}$.

Assumptions imposed for the estimation of these model parameters include the mean of the residuals equal to zero ($E(\varepsilon) = 0$), the residuals are uncorrelated with one another ($cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$), the residuals are uncorrelated with the latent factors ($cov(\varepsilon_i, \eta) = 0$), and that the model is linear in the parameters (e.g., all model parameters enter the model linearly; Bollen, 1989). Although the validity of these assumptions can be evaluated to varying degrees, the assumption of linearity is almost always violated in the presence of categorical data. In fact, the presence of categorical indicators implicitly introduces a nonlinear relationship into the system of equations. Two key assumptions are violated when treating an ordinal measure as if it were continuous.

First, we can think of each observed repeated ordinal measure in vector $\mathbf{y}$ linked to a corresponding continuous underlying measure $\mathbf{y}^*$. The standard linear model assumes that $\mathbf{y} = \mathbf{y}^*$; that is, that the repeated measures are continuous (or at least rough approximations). However, when we have observed a dichotomous or ordinal $\mathbf{y}$, this equality will not hold. In this situation, our fitted growth models will hold for $\mathbf{y}^*$ but *not* for $\mathbf{y}$. Second, we can write the moment structure hypotheses as $\boldsymbol{\Sigma}^* = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\mu}^* = \boldsymbol{\mu}(\boldsymbol{\theta})$, where $\boldsymbol{\Sigma}^*$ is the population covariance matrix of $\mathbf{y}^*$, $\boldsymbol{\mu}^*$ is the vector of means of $\mathbf{y}^*$, $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is

the implied covariance matrix, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the implied mean vector, and $\boldsymbol{\theta}$ is the vector of model parameters. Given that the observed variables $\mathbf{y}$ are collapsed versions of $\mathbf{y}^*$, in nearly all cases $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}^*$ and $\boldsymbol{\mu} \neq \boldsymbol{\mu}^*$ so that $\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{\mu} \neq \boldsymbol{\mu}(\boldsymbol{\theta})$. Thus, the moment structure hypotheses will typically not hold for the observed repeated ordinal measures.

Part of the corrective procedure to overcome these violations is to explicitly link $\mathbf{y}$ to $\mathbf{y}^*$(Jöreskog, 1994; Muthén, 1983, 1984; Muthén & Satorra, 1995). An auxiliary threshold model is a natural way to show the nonlinear relation between these categorical and continuous variables in which $y_{it} = c_t$ when $\tau_{c_t-1} < y_{it}^* \leq \tau_{c_t}$ where $c_t = 1, 2, 3, ..., C_t$ the total number of ordered categories, τ_{c_t-1}, τ_{c_t} are the lower and upper thresholds for category c_t with $\tau_{0_t} = -\infty$ and $\tau_{C_t} = +\infty$, and the thresholds are ordered from lowest to highest. This model connects the ordinal variable to its underlying continuous counterpart such that when the continuous variable lies between two thresholds, the ordinal variable will register in the category those thresholds determine. The mean ($\mu_{y_t^*}$) and variance ($\sigma_{y_t^* y_t^*}$) of y_{it}^* are unknown as are the thresholds from τ_{1t} to τ_{C_t-1}, but a subset of these can be estimated from the observed multivariate frequency distribution (Mehta et al., 2004). Finally, the correlation structure among the unobserved response distributions can be estimated, and these are often referred to as tetrachoric (for dichotomously scaled items) or polychoric (for polytomously scaled items) correlations.

The classic estimator in SEM for fitting polychoric correlation matrices is Weighted Least Squares (WLS) which is given as

$$F_{WLS} = [\widehat{\boldsymbol{\rho}}^* - \boldsymbol{\rho}^*(\boldsymbol{\theta})]' \mathbf{W}^{-1} [\widehat{\boldsymbol{\rho}}^* - \boldsymbol{\rho}^*(\boldsymbol{\theta})], \tag{10}$$

where $\widehat{\boldsymbol{\rho}}^*$ is a vector that contains all the estimated polychoric correlations and the means of the $\mathbf{y}^*$ variables, $\boldsymbol{\rho}^*(\boldsymbol{\theta})$ is the model-implied values of the polychoric correlations and the means, $\boldsymbol{\theta}$ contains all of the model parameters, and $\mathbf{W}$ is an optimal weight matrix (Browne, 1984). However, limitations of WLS include the need for very large sample sizes for the preceding asymptotic properties to hold, which has led researchers to seek alternative estimators including diagonal WLS (Jöreskog & Sörbom, 1979), corrected ML (Jöreskog, Sörbom, Du Toit, & Du Toit, 1999), robust WLS (Muthén, Du Toit, & Spisic, 1997), and 2SLS (Bollen, 1996), among others.

Recall that one of our goals is to use a two-stage process in which we first fit a measurement model to our observed categorical repeated measures, and then fit a growth model to the resulting individual specific scale scores. Equations 8 and 9 provide insight into a method for obtaining individual latent scores, also known as factor scores (e.g., Grice, 2001). Specifically, if we can estimate the relationship between the items and obtain an estimate of the individual variability

around the latent factor, we can work backwards to obtain the individual factor scores. Given that these factor scores take both the differentially weighted items as well as measurement error into account, they provide a more accurate estimate of individual scores and variability than the standard proportion score methodology.

The estimation of the CCFA model just described above provides estimates of the covariance among the latent factors, $\hat{\Psi}$, the factor loadings, $\hat{\Lambda}_y$, and the model-implied covariance matrix, $\hat{\Sigma}_{yy}$. These estimates can then be used to obtain the ordinary least squares regression coefficients used to premultiply y resulting in individual factor score estimates, $\hat{\eta}_i$.

Formally,

$$\hat{\eta} = \hat{\Psi}\hat{\Lambda}_y'\hat{\Sigma}_{yy}^{-1}\mathbf{y} \ , \tag{11}$$

where all parameters are as defined earlier. Once factor scores have been obtained via Equation 11, these can then be used within a second stage of analysis such as the standard multilevel framework.

SUMMARY

In sum, we have described three possible approaches for obtaining individual and time specific measures of an underlying construct based on a set of observed discretely scaled items: the proportion score methodology, the 2PL IRT, and the categorical CFA. Each approach is characterized by certain advantages and disadvantages. Of key importance to our goals here, despite rather widespread use in practice, the proportion score appears to be the most limited given the imposition of equal item weighting and the untestable assumption of measurement invariance. In contrast, both the IRT and CCFA models allow for differential item weighting as a function of severity and the strength of the relation between the item and the underlying construct. Furthermore, the IRT model provides formal tests of differential item functioning (e.g., Thissen, Steinberg, & Wainer, 1988) and the CCFA models offer the possibility of formal tests of measurement invariance (e.g., Meredith, 1964; 1993). This is particularly useful for evaluating the equivalence of measurement structures across different contexts. To better understand the applicability of these three approaches in developmental research, we now turn to an empirical examination of the relative behavior of these three methods when fitted to a single empirical sample of children followed from ages 10 to 21.

TRAJECTORIES OF EXTERNALIZING SYMPTOMATOLOGY IN ADOLESCENTS

The Adolescent and Family Development Project

To compare the three methods of scoring categorical measurement scales, we examined data drawn from the Adolescent and Family Development Project (AFDP: Chassin, Rogosch, & Barrera, 1991). The total sample consisted of 454 adolescents and their parents who completed repeated computerized, in-home interviews. Of these, 246 included a biological and custodial alcoholic parent whereas 208 were matched controls. COA families were recruited by means of court records (n = 103), wellness questionnaires from a health maintenance organization (n = 22), and community telephone surveys (n = 120). Inclusion criteria for COA families were Hispanic or non-Hispanic Caucasian ethnicity, Arizona residency, having a $10-15$ year old adolescent, English-speaking, and lack of cognitive limitations precluding an interview.

Lifetime presence of parent alcoholism was determined through diagnostic interviews with parents using the Diagnostic Interview Schedule or through spousal report using the Family History Research Diagnostic Criteria (if the alcoholic parent was not interviewed). Matched control families were recruited by phone screens of families identified through reverse directory searches based on identified COAs. Control families matched COA families on the basis of ethnicity, family composition, target child's age (within 1 year), and socioeconomic status (using the property value code from the reverse directory). Direct interview data confirmed that neither biological nor custodial parents met criteria for a lifetime alcoholism diagnosis.

These families were initially interviewed when the adolescents were aged $10-15$ (wave 1) and re-interviewed on an annual basis when the adolescents were aged $12-16$ (wave 2) and $13-17$ (wave 3), and again after a 5 year lag when the target adolescents were aged $18-21$ (Wave 4). Sample retention has been high, with 444 (97.8%) interviewed at all of the first three waves and 407 (90%) were interviewed again at the fourth wave. For the current analysis, 444 adolescents were considered with a complete age range of 10 to 21.[4]

The theoretical construct of interest here is adolescent externalizing symptomatology. For our demonstration, we considered a subset of eight items drawn from the Child Behavior Check List (CBCL; Achenbach & Edelbrock, 1983).

[4]There were 29 children who are 10 years of age at the first assessment. These subjects were included in the calibration and scoring steps, but were omitted from the multilevel growth models given the small sample size at the first observed age. All substantive findings were similar for all analyses when including these 29 cases.

TABLE 5.1
Item Parameters From Proportion Scores (p), 2PL IRT (a and b), CCFA (λ and τ), From Calibration Sample of 435 10 – 15 Year Old Children for the Eight CBCL Items at First Wave of Assessment

Item	*Item Wording*	p	a	b	λ	τ
1	Argues a lot	0.81	2.27	−1.13	0.79	−0.89
2	Cruelty, bullying, or meanness to others	0.24	2.12	0.92	0.76	0.70
3	Destroys things	0.06	1.42	2.52	0.56	1.61
4	Gets in many fights	0.17	2.66	1.14	0.81	0.94
5	Lying or cheating	0.29	1.29	0.89	0.64	0.54
6	Physically attacks people	0.07	1.81	2.04	0.64	1.45
7	Threatens people	0.15	2.73	1.26	0.83	1.04
8	Steals	0.07	1.53	2.20	0.64	1.45

Note. p = proportion of sample endorsing item, a = IRT discrimination, b = IRT severity, λ = CCFA factor loading, and τ = CCFA threshold.

These items were mother's report of the extent to which the adolescent argues, is cruel, destroys things, fights, lies, attacks others, threatens, or steals (see Table 5.1 for details). The data were originally collected using a three-response format (not true, somewhat or sometimes true, very true or often true), but there were sparse responses within and across time in the highest category denoted "very true or often true". Such sparseness introduces instability in the estimation of parameters in the subsequent measurement models (Thissen, Chen, & Bock, 2003). Because of this, the trichotomous response was collapsed to a binary response of "not true" and "sometimes or very true".

Calculation of Scale Scores

We used the three scoring methods described earlier to compute individual and time specific scale scores of externalizing symptomatology based on the eight CBCL items assessed on the 444 adolescents from the AFDP. For the IRT and CCFA approaches, we needed to use a calibration sample from which to estimate the parameters necessary for the scoring step. The sample used for the item calibration consisted of all children between age 10 and 15 at wave 1 of the mother's report of the eight CBCL items. This resulted in 435 cases (nine cases from the total sample did not meet the selection criteria).

Proportion Scores

Unlike the IRT and CCFA approaches, there is no calibration step necessary in the calculation of the proportion scores. Instead, the proportion scores were calculated using Equation 5 for adolescents within each individual age. For comparison purposes only, we present the proportion scores based on the calibration sample that was used for the IRT and CCFA models in Table 5.1. Thus, the column denoted p represents the proportion of the eight CBCL items endorsed by all subjects in the calibration sample between the ages of 10 and 15 at first wave of assessment. We present the mean proportion scores across all ages next.

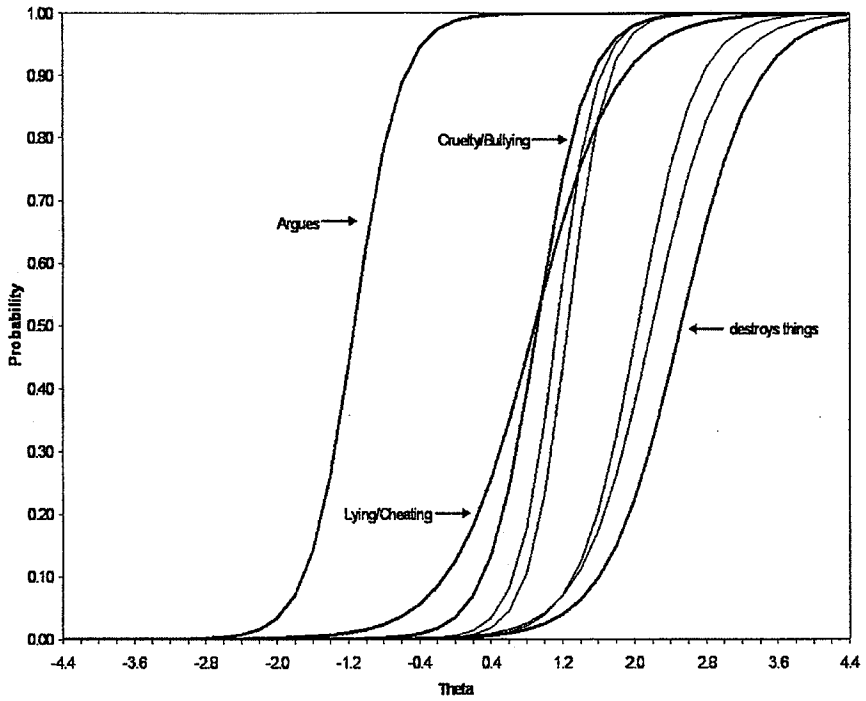
IRT Scores

The IRT analysis consisted of two steps: the calibration step and the scoring step. The calibration step provided the discrimination and severity parameters for each item (e.g., a and b from above), and the scoring step used these parameters to calculate the individual and time specific $\hat{\theta}$ values. We first fitted a standard 2PL IRT model to the eight dichotomously scored items on the 435 subjects in the calibration sample using maximum marginal likelihood estimation available in Multilog (Thissen, Chen, & Bock, 2003). This strategy allowed us to estimate the difficulty and discrimination parameters for each of the eight items that were then used in the creation of the individual scale scores.

The resulting IRT parameters are presented in Table 5.1, and the trace lines for eight items are presented in Figure 5.1 with four items specifically labeled ("argues a lot", "cruelty, bullying, or meanness", "lying or cheating", and "destroys things"). The discrimination parameter (i.e., a) reflects the steepness of the trace line for any given item. Comparing the "cruelty" item with the "lying" item shows the difference between slopes of 2.12 and 1.29, respectively. These numbers suggest that the "cruelty" item better discriminates at higher levels of externalizing symptomatology than does the "lying" item. Another useful analogy can be made between IRT discrimination parameters and factor loadings from the factor analysis literature. Discrimination parameters, much like factor loadings, can also be thought of as indexes of the extent to which an indicator is related to the latent construct (Takane & De Leeuw, 1987). Higher slopes indicate a stronger relationship between an item and the underlying construct. In the current example, we would say that the "cruelty" item is more strongly related to externalizing symptomatology than the "lying" item.

Comparing "argues" and "destroys things" shows the difference between items with severity values of −1.13 and 2.52, respectively. Following convention for setting the scale metric in IRT models, the b parameters are in a standard normal metric. This allows us to conclude that an individual would need to be approximately 1.1 standard deviations below the mean level of externalizing symptomatology to have a 50% chance of endorsing the "argues" item. On the

FIGURE 5.1
Item Response Theory trace lines for the eight CBCL items with "argues," "lying," "cruelty/bullying," and "destroys things" specifically labeled.



other hand, an individual would have to be approximately 2.5 standard deviations above the mean level of externalizing symptomatology to have a 50% chance of endorsing "destroys things". These values make intuitive sense: Individuals who destroys things are (generally) demonstrating greater externalizing symptomatology than individuals who argue. So, although arguing is still reflective of underlying externalizing symptomatology, endorsement of this item reflects a lower level of externalizing symptomatology than do other behaviors such as destroying things.

Finally, we used the item parameters from the calibration step to calculate individual- and time-specific scale scores for externalizing symptomatology for each of the 444 subjects. Multilog (Thissen et al., 2003) was used to produce modal a posteriori (MAP) scores for each subject at each time of assessment. The individual and time specific MAP is the mode of the posterior distribution of θ_{ti}, and we refer to these scores as $\hat{\theta}_{ti}$.

CCFA scores

As with the IRT approach, we again used a two step procedure to calculate the CCFA scores: a calibration step and a scoring step. For the calibration step, we used the same calibration sample from the IRT analysis. We estimated an eight item, one factor, CCFA model using robust weighted least squares (RWLS) fitted to tetrachoric correlations available in Mplus (version 2.14; Muthén & Muthén, 2001). The model was identified by constraining the latent mean to zero and latent variance to unity (see Figure 5.2 for a path diagram of this calibration model). The model fit the observed data well ($\chi^2\,(14) = 27.8$, $p = .015$, $CFI = .99$, $TLI = .98$, $RMSEA = .048$), and the final factor loadings (λ) and item thresholds (τ) are presented in Table 5.1.

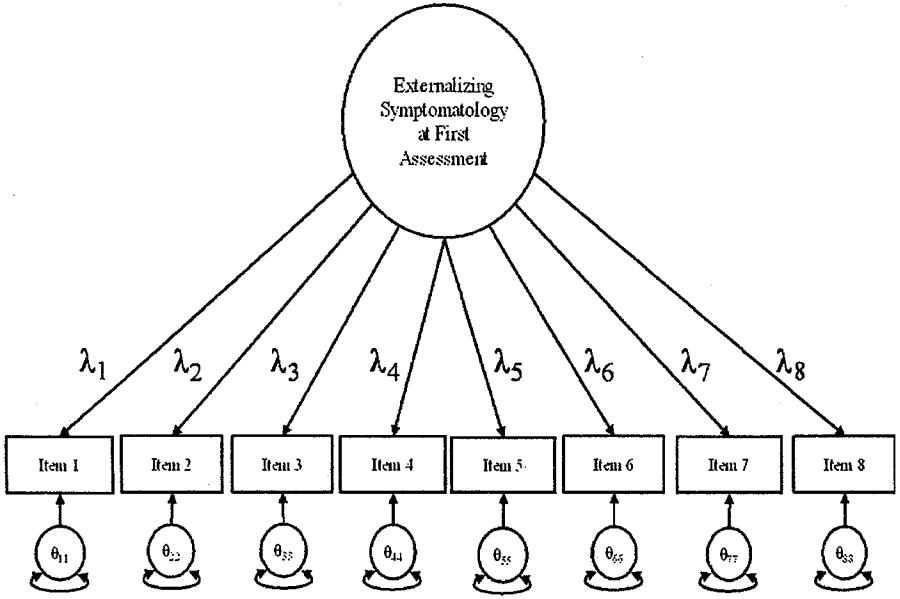
For the scoring step, these calibration parameter estimates were then used to calculate ordinary least squares factor-score regression coefficients (see Bollen, 1989, pp. 305 – 306)[5]. As part of the scoring, the observed binary data was recoded in order to take advantage of all available information. Specifically, endorsing an item was coded .5, failing to endorse an item was coded $-.5$, and missing responses were coded zero. This rescaling allowed all available information to be used in factor score estimation. The choice of $-.5$ and .5 retains the $1-$unit interval width between responses that was present in the 0/1 coding, yet allows us to incorporate the small number of cases with some subset of CBCL items missing within a given time period. Our motivation was thus to treat "missing" differently from "not endorsed" in the scoring step. This had the further advantage of maximizing the correspondence between the CCFA and IRT scoring methods. Using this approach, we calculated the individual and time specific factor scores based on the relevant model matrices in Equation 11. We refer to these estimated factor scores as $\hat{\eta}_{ti}$.

Calibration Invariance

When considering the study of development within and across context, it is critical that measurement models be evaluated for invariance across known contextual influences. For example, recall that our calibration sample consisted of children aged 10 to 15 at first assessment. This resulted in a mean age of 12.7, with 43% of the sample 12 years of age or younger. Given the importance of the IRT and CCFA parameters to the subsequent scoring steps that are based on this calibration sample, it is important to evaluate measurement invariance

[5] Although the methods described in Bollen (1989) were developed for continuously distributed indicators, we modified these calcluations slightly to allow for the incorporation of dichotomous indicators.

FIGURE 5.2
Path diagram for categorical CFA model fitted to the 8 CBCL items.



in terms of young (10, 11, or 12 years of age) versus old (13, 14, or 15 years of age) children. [6]

Unfortunately, the very nature of the psychometric model used for the calculation of the proportion scores does not allow for any testing of the invariance of this model as a function of child age. We can thus bring no empirical information to bear on this question for our estimates of p_{ti}.

To empirically evaluate the degree to which the item parameters from the 2PL IRT were invariant over age within the calibration sample, we performed differential item function (DIF) analyses using the freely available program IRTLRDIF (www.unc.edu/~dthissen/dl.html). Briefly, this approach allows for a series of one degree-of-freedom chi-square tests to ascertain the equality of the pair of parameters associated with each individual item (i.e., discrimination and severity) as a function of group membership. The results indicated that there was no significant DIF associated with any item parameter in comparing the "young" and "old" groups of children. This is strong evidence that the item

[6] It is important to remember that here we are evaluating the degree of measurement invariance as a function of young versus old children within the calibration sample only. This is thus testing invariance across children and over age, and is not a test of measurement invariance within child over age. This is an equally important issue to consider, a full treatment of which is beyond the scope of the current chapter.

parameters calculated on the full calibration sample and used in subsequent scoring do not vary for children aged 10 to 12 compared to children aged 13 to 15.

We were similarly able to empirically evaluate the invariance of the CCFA parameter estimates as a function of young versus old children using the multiple group approach as estimated in Mplus (Muthén & Muthén, 2001). The CCFA allows for an omnibus chi-square test of the equality of all parameter estimates across group membership. Two models were fitted to the two subsamples of young versus old children in the calibration sample. The first allowed all model parameters to be freely estimated over the two groups ($\chi^2\,(40) = 64.9, p = .008$, $CFI = .99$, $TLI = .98$, $RMSEA = .05$), and the second imposed equality constraints on all model parameters ($\chi^2\,(54) = 74.9$, $p = .02$, $CFI = .99$, $TLI = .99$, $RMSEA = .045$). Although the constrained model is nested within the unconstrained model, we can not calculate a formal chi-square difference test given our use of RWLS (Muthén et al., 1997). However, the constrained model reflects equal or superior fit on all indices compared to the unconstrained model thus suggesting that the equality constraints were appropriate relative to the characteristics of the sample data. These results converged with those of the IRT DIF analysis indicating that the measurement model of externalizing symptomatology was invariant with respect to age within the calibration sample.

Comparison of Three Scoring Methods

Following the methods described earlier, we calculated three different individual and time specific scale scores for the eight CBCL items: p_{ti}, $\hat{\theta}_{ti}$, and $\hat{\eta}_{ti}$. We compare these scores in three ways: in terms of means, correlations, and fixed and random effects from fitted multilevel growth models.[7]

Mean comparisons

The first comparison of interest relates to the pattern of change in age-specific means of externalizing in the three different score types over time. We present the age specific means of the externalizing symptomatology scale scores at each age for all three scoring methods in Table 5.2. Direct comparison of both the means and standard deviations are difficult given that all three scale scores are on different metrics (e.g., the proportion is bounded between 0 and 1, and the

[7] It is possible to draw more specific analytic relations between p_{ti}, $\hat{\theta}_{ti}$, and $\hat{\eta}_{ti}$, although a detailed treatment of this is beyond the scope of our chapter. See chapters 15 and 16 of Lord and Novick (1968) and Takane and de Leeuw (1987) for further details.

IRT and CCFA are theoretically bounded between negative and positive infinity). Furthermore, even within the IRT and CCFA, the underlying metrics are not comparable in absolute terms but only in relative terms. Thus, it is evident that the highest level of externalizing symptomatology across all three scale scores occurs at the youngest age of 11. Additionally, the externalizing symptomatology scores systematically decrease with increasing age. (Again, recall that through our use of the accelerated longitudinal design, we are able to track these scores over 11 distinct ages although no single child provided more than four waves of data.)

We can also compare these age-specific means graphically through the use of multiple $y-$axes to denote the scale of each score. In Figure 5.3 we present the pattern of means of the IRT and proportion scores from age 11 to 21.[8] The IRT scores are scaled on the left-hand y-axis, and the proportion scores are scaled on the right-hand y-axis. It is graphically clear that the age specific mean externalizing scores are both decreasing in a nonlinear trend for both the proportion and IRT scores. Note that these trends are quite similar from age 11 through 17, but begin to diverge slightly up to age 21. Although the trend lines are slightly offset, both scores appear to be approximately following the same trend over time.

Despite the similarity in age specific mean scores over time, this comparison does not allow us to consider characteristics of the individual scores within each age. To explore this, we next turn to an examination of the within time correlations among the three scores.

Correlation Comparisons

An important initial comparison is to simply assess the degree to which the three types of scores correlate with one another within each distinct age. We thus computed standard Pearson correlations among the three types of scores. Overall, the correlations among the three scores within a given age were extremely high. For example, across all ages the range of correlations between the proportion and IRT scores was $.96 - .97$, between the proportion and CCFA scores was $.97 - .99$, and between the IRT and CCFA scores was $.97 - .98$. Given that a squared correlation represents the proportion of overlapping variance between the two variables, the overall shared variability was $92\% - 98\%$. Clearly, there is a large amount of shared information among these three scoring methods.

[8]We do not present the CCFA scores here given that we can only use two different scales on the y-axis. However, the pattern of the CCFA scores nearly identically track that of the IRT scores.

TABLE 5.2
Age Specific Means (and Standard Deviations) for the Proportion (p), IRT ($\hat{\theta}$), and CCFA ($\hat{\eta}$) Scale Scores of Externalizing Symptomatology

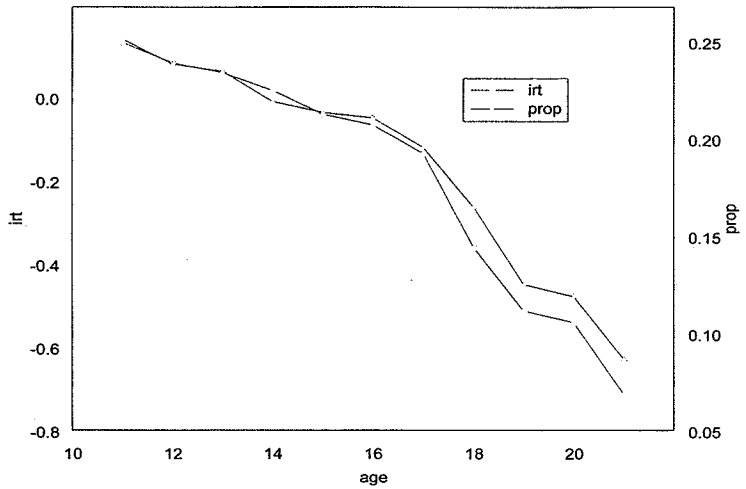
Age	*Sample Size*	p	$\hat{\theta}$	$\hat{\eta}$
11	103	.25(.21)	.13(.82)	−.26(.27)
12	181	.24(.20)	.08(.81)	−.28(.27)
13	256	.23(.20)	.06(.84)	−.28(.28)
14	281	.22(.19)	.02(.79)	−.30(.26)
15	236	.21(.20)	−.04(.82)	−.31(.27)
16	143	.21(.21)	−.07(.85)	−.31(.28)
17	55	.20(.21)	−.13(.88)	−.34(.28)
18	76	.16(.25)	−.36(.97)	−.38(.31)
19	78	.13(.18)	−.51(.82)	−.44(.24)
20	95	.12(.18)	−.54(.88)	−.44(.25)
21	91	.09(.14)	−.71(.72)	−.49(.20)

However, these high correlations do not yet allow us to conclude that these these three scoring methods offer an equivalent representation of the underlying construct. The reason is that the correlations reflect a high degree of overlap in the rank ordering of scores relative to the score mean; yet these do not reflect potentially important differences in other aspects of the score characteristics.

Most importantly, the correlations do not reflect potential differences in the variability among the individual scores within and across time. Given that the IRT and CCFA scores incorporate differential item severity, but the proportion scores do not, we predicted that there would be potentially important differences with respect to individual variability among the scores. This is particularly salient given our desire to fit random coefficients growth models to these scores, a type of analysis in which variability plays a key role.

To better understand this issue, consider the values presented in Table 5.3. The first eight columns reflect different patterns of endorsement of the eight items just within age 14. Consistent with how the proportion scores are calculated, the seven unique combinations of endorsement of any two items observed in the sample resulted in a proportion score of .25. There is thus no variability within the proportions scores for the endorsement of 2 of 8 items at age 14. However, recall that the IRT and CCFA scoring approach explicitly incorporate information about differential item properties. This means that the associated IRT or CCFA score for the endorsement of 2 of 8 items depends on *which* two items were endorsed. This in turn introduces greater variability among the individual and time specific scores, variability that might play a critical role when fitting random coefficient growth curve models.

FIGURE 5.3
Age-specific means of proportion and IRT scores for 8 CBCL items with IRT scores scaled on the left y—axis, and proportion scores scaled on the right.



Note. IRT represents item response theory score and "prop" represents proportion score.

Multilevel Growth Curve Models

As a final comparison of the three methods of scoring the eight dichotomous items, we fitted a random coefficients multilevel growth model using SAS PROC MIXED (SAS Institute, 1999). Importantly, we used chronological age and not wave of assessment as the time metric of interest. There were thus 11 distinct ages ranging from 11 to 21, although no single child was assessed more than four times (see, e.g., Mehta & West, 2000). We began by fitting fixed and random effects for a linear trajectory as a function of chronological age. Age was scaled such that 0 represented age 11; this allowed for the intercept term to be interpreted as the model-implied mean externalizing at the youngest age in the sample. This model was fit to all three scores separately, and significant fixed and random effects were found for the intercept and linear slope term in all three models.

However, as is evidenced from the mean trends presented in Figure 5.3, it seemed likely that a curvilinear trajectory might better capture the observed pattern of change over time. Thus a quadratic term was added to each of the three models, and the addition of the squared term of age significantly improved model fit as tested by the likelihood ratio test (e.g., testing the difference in deviance statistics between the more and the less restricted models; Raudenbush

TABLE 5.3
Proportion Scores (p), IRT Scores ($\hat{\theta}$), CCFA Scores ($\hat{\eta}$), and Frequency of Endorsement ($freq$) of any 2 of the 8 CBCL Items Observed in the Sample at Age 14

1	2	3	4	5	6	7	8	$freq$	p	$\hat{\theta}$	$\hat{\eta}$
1	0	0	0	1	0	0	0	19	.25	.175	−.312
1	0	1	0	0	0	0	0	1	.25	.214	−.336
1	0	0	0	0	1	0	0	4	.25	.322	−.308
1	1	0	0	0	0	0	0	18	.25	.401	−.241
0	1	0	0	0	0	1	0	1	.25	.507	−.185
1	0	0	1	0	0	0	0	7	.25	.524	−.194
1	0	0	0	0	0	1	0	6	.25	.539	−.159

Note. The integers 1 through 8 represent the CBCL item presented in Table 5.1, and $freq$ is frequency of endorsement of those two items at age 14.

& Bryk, 2002). We selected the quadratic trajectory model for interpretation here.

The point estimates and standard errors for the fixed and random effects of the quadratic model fitted to the three different scores are presented in Table 5.4. The fixed effects for all three scores are rather similar. That is, although the point estimates differ depending on scoring method (as would be expected given the differences in scaling described earlier), the pattern of significance is quite similar. Specifically, there is a significant fixed effect for the intercept, a nonsignificant fixed effect for the linear slope, and a significant fixed effect for the quadratic slope. This reflects the general trend seen in the mean trends in Figure 5.3 and in the model-implied mean trajectory for the IRT scores presented in Figure 5.4. There is a decreasing trend in the means over time, but the rate of decrease is larger as time progresses. Thus, in terms of the fixed effects, the substantive conclusions are quite similar across the three scoring methods.

However, these results do not reflect potential differences introduced into the multilevel model due to increased variability in the IRT and CCFA scores. Consistent with this prediction, there are important differences among these scoring methods in terms of the random effects of the trajectories that define the quadratic trajectory. Although there is a significant random effect for the intercept term for all three score types, differences across these models were evident for the random slope effects. For the proportion score, the random effect for the linear term is only marginally significant ($p = .074$) and the random effect for the quadratic term is nonsignificant ($p = .143$). However, the random effect for the linear term is significant for both the CCFA ($p = .030$) and IRT ($p = .013$) models. Moreover, the quadratic term is marginally signi-

TABLE 5.4
Fixed and Random Effects for Quadratic Growth Curve Model Fitted to all Three Scale Scores

Parameter	p	$\hat{\theta}$	$\hat{\eta}$
$\hat{\gamma}_{00}$	.246 (.014)*	.117 (.057)*	−.268(.018)*
$\hat{\gamma}_{10}$	−.003 (.006)^{ns}	−.008 (.024)^{ns}	−.002(.008)^{ns}
$\hat{\gamma}_{20}$	−.001 (.0005)*	−.008 (.002)*	−.002(.0007)*
$\hat{\tau}_{00}$	.025 (.004)*	.447 (.078)*	.046(.008)*
$\hat{\tau}_{11}$	.001 (.001)^{ns}	.040 (.018)*	.003(.001)*
$\hat{\tau}_{22}$	.0001 (.0001)^{ns}	−.003 (.0001)*	.00002(.00001)§
$\hat{\sigma}^2$	.014 (.001)*	.238 (.013)*	.026(.001)*

Note. $\hat{\gamma}_{00}$ = fixed intercept; $\hat{\gamma}_{10}$ = fixed linear; $\hat{\gamma}_{20}$ =fixed quadratic; $\hat{\tau}_{00}$ =random intercept; $\hat{\tau}_{11}$ =random linear; $\hat{\tau}_{22}$ =random quadratic; $\hat{\sigma}^2$ =level-1 residual; $ns = p > .05$; $* = p < .05$; § $= p < .10$.

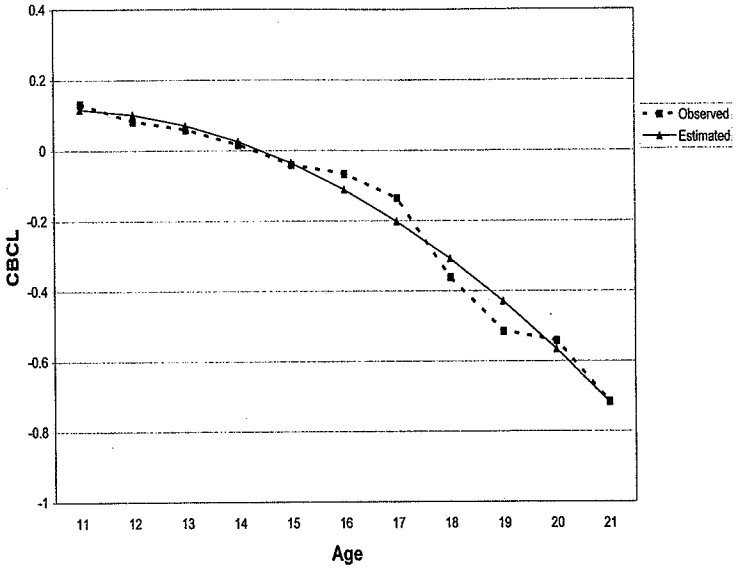
ficant for the CCFA scores ($p = .072$), and is significant for the IRT scores ($p = .031$). As predicted, the added variability associated with CCFA and IRT scoring procedures is reflected in greater individual variability in the growth curve models.

These differences could be quite salient from a substantive point of view. Fitting the growth model to the proportion scores would suggest that although the group curve of externalizing is decreasing nonlinearly over time, there is no evidence of individual variability in rates of change around this mean curve. In comparison, the CCFA and, even more clearly, the IRT results suggest that there is meaningful individual variability among the externalizing trajectories over time. This supports a very different substantive conclusion than that derived from the analysis of the proportion scores and demonstrates the impact that differing scoring procedures can have on evaluating change over time.

CONCLUSIONS

A major part of what makes empirical research in developmental psychology both fun and rewarding is the very thing that makes it an equally vexing task. Namely, it is hard; and it is even harder to do well. Consider just a random sampling of issues that commonly arise in the study of developmental processes over time: missing data, categorical and non-normally distributed repeated measures, nonlinear trajectories, nesting of children within families or schools, changes in the manifestation or even meaning of a construct over development. Furthermore, consistent with the theme of this book, all of these issues are even more salient when attempting to study the development of individuals

FIGURE 5.4
Observed means (dashed line) and model-implied quadratic trajectory (solid line) for IRT scores from the multilevel growth model fitted to the sample of 439 adolescents.



embedded within one or more contexts. Finally, these issues all assume that the data has already been collected; equally challenging issues arise in the design and execution of a developmental study.

Among these many challenges, we have focused our attention on the incorporation of psychometric scales made up of multiple dichotomous items in the analysis of individual growth curves over time. We described a two-stage procedure in which scale scores are first calculated, and growth curve models are then fitted to these scores. In contrast to this approach, there exist several advanced analytic methods that allow for the simultaneous estimation of a formal measurement model of a set of dichotomous items with a growth curve model. Although unquestionably the ideal strategy, these techniques impose certain restrictions that may not be appropriate in many areas of developmental research and, more importantly, these methods are also currently difficult (if not impossible) to implement in practice. Given our desire to consider methods that can be applied in developmental research settings, we have approached this problem using the two-stage strategy.

For the first analytic stage we considered three methods for calculating the scale scores based on a set of dichotomous items: the proportion score, the 2PL IRT model, and the CCFA model. Not surprisingly, the resulting scores from all

three of these methods were highly correlated with each other. Indeed, within age correlations ranged from .96 – .99 with most in the .98 range. This prompts one to wonder if all of the added complexity of the IRT and CCFA approaches are worthwhile given these high linear relations. However, we strongly believe that these alternative approaches are important to consider because these high correlations only tell part of the story.

More specifically, the proportion score simply reflects the number of dichotomous items that were endorsed by an individual within an assessment period. Thus a score of .25 unambiguously reflects that two of the eight items were endorsed. However, what is critical to understand is that this same score of .25 results from the endorsement of *any* two of the eight items. So endorsing "lying" and "arguing" is assigned the same numerical value as endorsing "cruelty" and "physically attacks people". It may be (and our findings suggest that it is) very important to use a measurement model that allows for differential contributions of these items depending upon severity and the strength of the relation of the item to the underlying construct of externalizing symptomatology. For this, we turned to the IRT and CCFA models.

In contrast to the proportion score, the IRT and CCFA methods explicitly allow for differential item weighting in the creation of a scale score. Importantly, it is not just that two of eight items were endorsed, but *which* two of the items were endorsed. Because different scores are assigned based on the specific set of items that were endorsed, this allows for greater variability among the individual and time specific scores. This increased variability was clearly evident when comparing the different IRT and CCFA scores resulting from different patterns of item endorsement (see Table 5.3). This in turn increases the variability available for modeling in the subsequent growth curve analysis. Whereas the results of the growth model fitted to the proportion scores indicated no individual variability in developmental trajectories over time, the very same model fitted to the IRT and CCFA scores indicated significant individual variability. This would have been missed had we relied solely on the proportion scores alone.

Potential Model Extensions

We have only touched on a few of the potential applications of these methods for testing a variety of research hypotheses derived from developmental theory. Specifically, we considered a rather simple situation in which there was a single set of eight dichotomous items assessed at each time period. However, the IRT and CCFA methods naturally extend to more complex and more interesting conditions.

For example, the full AFDP study incorporated 25 items to assess externalizing symptomatology in the first three waves of study, but only 8 of the 25 items in the fourth wave. We used the common 8 items here, but these models could be expanded to include all available items within any given wave of assessment. This would not only allow for greater precision in measuring the underlying construct, but would also allow for changing item sets over time that were calibrated on the same metric to allow for subsequent growth modeling. A particular advantage of this approach is the possibility for addressing heterotypic continuity; that is, developmental theory often predicts that the very manifestation of a theoretical construct changes across context and development. The explicit modeling of changing item sets across development is one method for addressing this challenge.

A related extension is to consider not only changing item sets over time within a study, but also changing item sets in data drawn from multiple studies. For example, we are currently applying these techniques to combine data from one study that spans ages 3 to 15 with a second study that spans ages 10 to 21. Using methods of item equating in IRT and multiple group equality constraints in CCFA, it is possible to combine data drawn from two or more independent data sets and to calculate scale scores that are on a comparable metric across study and over time. This allows for the study of trajectories from age 3 to 21, and is also a highly efficient use of existing data.

Additionally, although we focused on dichotomously scored items here, both the IRT and CCFA naturally extend to consider three or more response scales (e.g., trichotomous, ordinal, Likert, and so forth). These more complicated models do of course require more parameters to be estimated relative to the dichotomous outcome. However, given sufficient sample size, all of our IRT and CCFA developments presented here extend naturally to items with more than two response categories.

Finally, a particularly exciting future direction involves the implementation of new studies that are explicitly designed to include these methodological techniques. For example, a single study could be designed to include a set of shared "anchor" items over time, and subsets of items could rotate in and out depending on developmental relevance. Further more, the inclusion of a set of anchor items across multiple studies would expand the possibilities for item equating across study and over time. Combined with advances in accelerated longitudinal designs (Duncan, Duncan, & Hops, 1994) and planned missingness (e.g., Schafer & Graham, 2002), the inclusion of more comprehensive psychometric models can result in powerful and efficient experimental designs for the study of individual development across the lifespan.

Conclusions and Recommendations for Applied Researchers

As quantitative psychologists, we strive to strike a balance between what is ideal from a statistical perspective and what is feasible from a pragmatic perspective. Clearly the single-stage analytic strategy is the gold standard to which we aspire, but these methods are not currently applicable in many settings commonly encountered in developmental research. In comparison, we believe the two-stage procedures we used here allow for the inclusion of a more general psychometric model of our theoretical construct while retaining our ability to fit these models to actual data. We conclude with a few recommendations for the use of these techniques in developmental research.

Regarding the use of proportion scores, we see only a small number of situations where this is the optimal scoring strategy. As we have demonstrated both analytically and empirically, the proportion score is the least psychometrically sound approach of all three methods considered here. All items are unit-weighted in which information about item severity, discrimination, and reliability is disregarded. The situation in which the use of a proportion score might be beneficial is when the available sample size is too small to support the estimation of the more complex IRT or CCFA approaches. If nothing else, it would be beneficial to compare the proportion score analysis with those of the IRT or CCFA to determine if potentially valuable information is being lost using the former technique.

We see a number of significant advantages in the use of the IRT and CCFA scale scores. First, both approaches not only consider how many items were endorsed, but which items were endorsed; this allows for greater psychometric rigor and increased variability among the individual and time specific scores. Second, these methods allow for the inclusion of changing item sets over time and within study, or even across multiple studies. Third, both the IRT and CCFA approaches can provide formal tests of measurement invariance both in terms of the calibration step (as we demonstrated above), but also of longitudinal measurement invariance across development (which we did not pursue here). As we noted earlier, tests of invariance allow for powerful insights into measurement properties both within and across developmental contexts. Finally, the use of the IRT and CCFA methods naturally allow for access to all of the strengths of each of these approaches in isolation (e.g., DIF analysis with the IRT; multiple group analysis with the CCFA). Taken together, we believe these methods have much to offer the analysis of individual differences in developmental stability and change.

REFERENCES

Achenbach, T., & Edelbrock, C. (1983). *Manual for the child behavior checklist and revised child behavior profile.* Burlington, VT: University Associates in Psychiatry.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*, 134-167.

Biesanz, J. C., Deeb-Sossa, N. P., Aubrecht, A. M., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods, 9*, 30-52.

Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Bollen, K. A. (1996). An alternative 2SLS estimator for latent variable models. *Psychometrika, 61*, 109-21.

Bollen, K. A., & Curran, P. J. (2006). *Structural equations with latent variables. wiley series in probability and mathematical statistics.* New York: Wiley.

Bond, T. G., & Fox, C. M. (2001). *Applying the rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 127-141.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.

Chassin, L., Rogosch, F. A., & Barrera, M. (1991). Substance use and symptomatology among adolescent children of alcoholics. *Journal of Abnormal Psychology, 100*, 449-463.

Cicchetti, D., & Rogosch, F. A. (1996). Equifinality and multinality in developmental psychopathology. *Development and Psychopathology, 8*, 597-600.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*, 529-569.

Curran, P. J., & Willoughby, M. J. (2003). Implications of latent trajectory models for the study of developmental psychopathology. *Development and Psychopathology, 15*, 581-612.

Curran, P. J., & Wirth, R. J. (2004). Inter-individual differences in intraindividual variation: Balancing internal and external validity. *Measurement: Interdisciplinary Research and Perspectives, 2*, 219-227.

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data.* Boca Raton, FL: CRC Press.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford, England: Oxford University Press.

Duncan, T. E., Duncan, S. C., & Hops, H. (1994). The effect of family cohesiveness and peer encouragement on the development of adolescent alcohol use: A cohort-sequential approach to the analysis of longitudinal data. *Journal of Studies on Alcohol, 55*, 588-599.

Fox, J.-P. (2003). Stochastic em for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology, 56*, 65-81.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical Psychology*, 145-172.

Gibbons, R. D., & Hedeker, D. (1997). Random-effects probit and logistic regression models for three-level data. *Biometrics, 53*, 1527-1537.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

Gottlieb, G., & Halpern, C. T. (2002). A relational view of causality in normal and abnormal development. *Development and Psychopathology, 14*, 421-435.

Gottlieb, G., Wahlsten, D., & Lickliter, R. (1998). The significance of biology for human development: A developmental psychobiological systems view. In R. M. Lerner & D. William (Eds.), *Handbook of child psychology: Volume 1: Theoretical models of human development* (5th ed., pp. 233-273). New York: Wiley.

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

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika, 59*, 381-389.

Jöreskog, K. G., & Sorbom, D. (1979). *Advances in factor analysis and structural equation models.* Cambridge, MA: Abt Books.

Jöreskog, K. G., Sorbom, D., Du Toit, S., & Du Toit, M. (1999). *Lisrel 8: New statistical features.* Chicago, IL: Scientific Software.

Kagan, J. (1971). *Change and continuity in infancy.* Oxford, England: Wiley.

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

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology 1983* (pp. 72-103). San Francisco: Jossey-Bass.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed.). New York: Plenum Press.

McArdle, J. J. (1989). Structural modeling experiments using multiple growth functions. In P. Ackerman, R. Kanfer, & R. Cudeck (Eds.), *Learning and individual differences : Abilities, motivation and methodology* (pp. 71-117). Hillsdale, NJ: Lawrence Erlbaum Associates.

McCullagh, P., & Nelder., J. A. (1989). *Generalized linear models.* London, England: Chapman and Hall.

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods, 9*, 301-333.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177-186.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

Meredith, W., & Tisak, J. (1984). "Tuckerizing" *curves.* Santa Barbara, CA: Paper presented at the annual meeting of the Psychometric Society.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674-701.

Muthèn, B. O. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22*, 48-65.

Muthèn, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthèn, B. O. (1996). Growth modeling with binary responses. In A. V. Eye & C. Clogg (Eds.), *Categorical variables in developmental research: Methods of analysis* (pp. 37-54). San Diego, CA: Academic Press.

Muthèn, B. O., Du Toit, S., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Unpublished manuscript,* University of California, Los Angles.

Muthèn, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor-analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-180.

Muthèn, B. O., & Satorra, A. (1995). Technical aspects of muthèn's liscomp approach to estimation of latent variable relations with a comprehensive measurement model. *Psychometrika, 60*, 489-503.

Muthèn, L. K., & Muthèn, B. O. (2001). *Mplus user's guide.* Los Angeles: Muthèn & Muthèn.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 35-64). Washington, D.C: American Psychological Association.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models. applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage Inc.

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A multivariate multilevel rasch model with application to self-reported criminal behavior. *Sociological Methodology, 33*, 169-212.

Sameroff, A. J. (1995). General systems theories and developmental psychopathology. In D. J. Cohen & D. cicchetti (Eds.), *Developmental psychopathology, vol. 1: Theory and methods* (pp. 659-695). Oxford, England: Wiley.

SAS Institute, I. (1999). *Sas documentation, version 8.* Cary, NC: SAS Publications.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*, 147-177.

Thissen, D., Chen, W.-H., & Bock, R. (2003). *Multilog (version 7) [computer sotware].* Lincolnwood, IL: Scientific Software International.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the

study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 659-695). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Lawrence Earlbaum Associates.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363-381.

Wirth, R. J., & Edwards, M. C. (2005). *Recent advances in item factor analysis. under review.*

Wohlwill, J. F. (1970). The age variable in psychological research. *Psychological Review, 77*, 49-64.

Wohlwill, J. F. (1973). *The study of behavioral development.* New York: Academic Press.

Y., T., & Leeuw J. de. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics, 44*, 1049-1060.

CHAPTER SIX

Representing Contextual Effects in Multiple-Group MACS Models

Todd D. Little
University of Kansas

Noel A. Card
University of Arizona

David W. Slegers
Emily C. Ledford
University of Kansas

In keeping with the goals of this volume, we explore the various uses and advantages of mean and covariance structures (MACS) models for examining the effects of ecological/contextual influences in developmental research. After addressing critical measurement and estimation issues in MACS modeling, we discuss their uses in two general ways. First, we focus primarily on discrete ecological factors as grouping factors for examining main-effect as well as moderating or interactive influences. Second, we briefly discuss the simplest case — including ecological factors as within-group direct and mediated effects — because these types of effects are covered in more detail elsewhere in this volume (see Little, Card, Bovaird, Preacher, & Crandell, chap. 9, this volume; see also McKinnon, in press). Our focus in this chapter will be on how such effects might be moderated by the discrete contextual factor(s) used to define groups.

Discrete ecological factors can be conceptualized at various levels, from macrosystems such as sociocultural contexts to exosystem structures such as neighborhoods and communities. Other discrete ecological factors such as developmental level, ethnicity, and gender are particularly amenable to study using