Structural Equation Modeling, 15:676–704, 2008

ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705510802339080

Psychology Press
Taylor & Francis Group

TEACHER'S CORNER

Incorporating Measurement Nonequivalence in a Cross-Study Latent Growth Curve Analysis

David B. Flora
York University

Patrick J. Curran and Andrea M. Hussong
University of North Carolina at Chapel Hill

Michael C. Edwards
The Ohio State University

A large literature emphasizes the importance of testing for measurement equivalence in scales that may be used as observed variables in structural equation modeling applications. When the same construct is measured across more than one developmental period, as in a longitudinal study, it can be especially critical to establish measurement equivalence, or invariance, across the developmental periods. Similarly, when data from more than one study are combined into a single analysis, it is again important to assess measurement equivalence across the data sources. Yet, how to incorporate nonequivalence when it is discovered is not well described for applied researchers. Here, we present an item response theory approach that can be used to create scale scores from measures while explicitly accounting for nonequivalence. We demonstrate these methods in the context of a latent curve analysis in which data from two separate studies are combined to estimate a single longitudinal model spanning several developmental periods.

Correspondence should be addressed to David B. Flora, 322 BSB, Department of Psychology, York University, 4700 Keele Street, Toronto, ON M3J 1P3 Canada. E-mail: dflora@yorku.ca

Empirically evaluating longitudinal trajectories of a construct over an extended period of time is associated with a host of complexities. These complexities include missing data that are both planned (e.g., due to an accelerated design) and unplanned (e.g., due to subject attrition), and potential measurement nonequivalence due to developmental period or other factors. By combining data sets from two or more existing longitudinal studies, researchers may be able to consider a longer period of development than is covered by any single study while also alleviating within-study sample size limitations due to missing data. This approach has been termed *mega-analysis* (McArdle & Horn, 2002) or *cross-study analysis* (Hussong, Flora, Curran, Chassin, & Zucker, 2008).

Prior to fitting structural equation latent curve models to repeated observations, it is essential to establish the equivalence, or invariance, of measurement structures over time (e.g., Bollen & Curran, 2006; Khoo, West, Wu, & Kwok, 2006). For example, endorsement of an item about crying behavior may be more strongly indicative of internalizing symptomatology for adolescents than for younger children, among whom crying may be more normative. If this source of nonequivalence is ignored, younger participants may be given spuriously higher scores on an internalizing scale. By having artificially higher scores at younger ages, the estimated longitudinal change in internalizing from childhood to adolescence can be biased relative to the true change. In a cross-study analysis, the importance of measurement equivalence is amplified because of the need to establish invariance across the separate studies contributing the longitudinal data. In this situation, certain characteristics of the sampling schemes of the contributing studies may lead to different measurement properties that impact subsequent conclusions drawn from the combined data set.

Although methods for testing measurement equivalence are well documented (e.g., Reise, Widaman, & Pugh, 1993), procedures for dealing with nonequivalence when it is found are not well described for applied researchers. Thus, the primary goal for this article is to describe the use of methods drawing from item response theory (IRT) that may be employed to create scale scores that explicitly account for measurement nonequivalence. Use of such scores, relative to standard scoring methods ignoring nonequivalence, leads to improved validity of subsequent structural equation modeling (SEM), such as latent curve analyses. Additionally, we describe similarities (and differences) between the IRT approach and methods relying on confirmatory factor analysis (CFA). Therefore, the purpose of this article is not to present new analytical developments, but rather to show a detailed example of how to account for measurement nonequivalence in practice.

Furthermore, this article demonstrates how these methods may be applied in a cross-study analysis, where data from more than one study are combined to estimate a single model. We present the approach for incorporating measurement nonequivalence in the context of a longitudinal cross-study model initially

presented by Hussong et al. (2008) examining internalizing symptomatology from early childhood to late adolescence. Here, in addition to providing detailed discussion of the IRT scoring procedure, we expand on those analyses by comparing latent curve model results that account for measurement nonequivalence with results that ignore measurement nonequivalence. In so doing, we provide detailed discussion about when measurement nonequivalence is likely to influence subsequent structural equation analyses.

ITEM RESPONSE THEORY

In structural equation growth modeling applications, repeated measures of an outcome construct are commonly created from multi-item scales by calculating the sum or mean of item responses within a given time period, with the items themselves typically producing dichotomous or ordinal distributions of responses. The sum- and mean-score methods lead to values that are a simple linear transformation of each other, and, if the items are dichotomous, the proportion of endorsed items. However, these methods cannot account for potential measurement nonequivalence in any straightforward fashion. IRT provides a powerful alternative methodology to basic sum scoring that can be particularly useful in longitudinal analysis (Curran, Edwards, Wirth, Hussong, & Chassin, 2007; Khoo et al., 2006; Seltzer, Frank, & Bryk, 1994). In particular, Curran et al. (2007) showed that the use of IRT-scaled scores, relative to simple proportion scores, leads to greater individual variability in the observed scores that can be used in subsequent analyses, such as a latent curve analysis. This additional variability then has implications for finding statistically significant model parameters, such as the variance of latent growth factors. Furthermore, of key importance for this article is that IRT is readily extended to situations in which there is measurement nonequivalence due to age differences or other covariates.

IRT encompasses a class of measurement models for categorical item-level data with the purpose of estimating parameters that describe the relation between each item and a latent construct (see Embretson & Reise, 2000). A crucial advantage of IRT is the ability to create scores on a common metric across more than one experimental design (e.g., a cross-study analysis where two different studies administer the same or similar instruments). Another advantage of IRT is that, unlike sum- or mean-score approaches, items with stronger relations to the latent construct are given more weight in scoring (i.e., through the discrimination parameter). Additionally, by ordering items according to their severity parameters, IRT-scaled scores, relative to sum or mean scores, more closely approximate interval-level measurement, which is crucial for latent curve modeling (see Khoo et al., 2006). Finally, as mentioned earlier, IRT easily

incorporates techniques for evaluating measurement equivalence across discrete groups using tests of differential item functioning (DIF; e.g., Thissen, Steinberg, & Wainer, 1993).

There are relatively few discussions available to applied researchers of how to account for DIF explicitly in subsequent modeling analyses. Therefore, in this article, we describe and demonstrate the incorporation of DIF for creating IRT-scaled scores for a measure of child and adolescent internalizing symptomatology. We then use these scores as observed variables in a latent curve model of internalizing from age 2 to 17. Furthermore, we show how this IRT approach facilitates a cross-study analysis in which we use DIF testing to establish a common scale of measurement across two longitudinal studies and then fit a latent curve model to the combined data from both studies. In addition to describing how to account for DIF, an important aspect of this article is comparing IRT-scaled scores and subsequent growth model results that do and do not incorporate DIF. In so doing, we examine DIF effect size to help illuminate whether DIF is likely to have an effect on ensuing analyses using IRT-scaled scores.

Two-Parameter Logistic Item Response Model

All IRT analyses presented here use the two-parameter logistic (2PL) model for dichotomous items (Birnbaum, 1968). This model uses a logistic function, often called a trace line or item characteristic curve, to describe the probability of endorsing a given item j as

$$P(y_j = 1|\theta) = \frac{1}{1 + \exp[-1.7a_j(\theta - b_j)]}, \quad (1)$$

where y_j is the observed response, a_j is the discrimination parameter, and b_j is the severity parameter. The continuous latent construct measured by the set of items is represented by θ. In the following analyses, θ is internalizing symptomatology as measured by the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1983). For scale identification purposes, θ is usually assumed to follow a normal distribution with a mean of zero and variance equal to one. The discrimination (or slope) parameter describes the extent to which an item is related to the latent construct, and the severity (or location) parameter defines the point along the latent continuum at which there is a 50% probability of item endorsement (the value of the severity parameter is also the inflection point of the logistic curve for a given item). Important assumptions for the 2PL model are that the set of items is unidimensional (i.e., there is a single latent construct, θ, accounting for the interrelations among items) and that the items are "locally independent," meaning that responses to a given item are completely independent of other responses when controlling for θ. Although this article illustrates

methods for dichotomous item response data, these are readily generalized to ordinal item response data using Samejima's (1969) graded response model, of which the 2PL model is a special case (Thissen, Nelson, Rosa, & McLeod, 2001). Figure 1 provides an illustration of 2PL trace lines, estimated in the analyses described later, for a single CBCL item that vary as a function of both age and gender, thus showing measurement nonequivalence, or DIF.

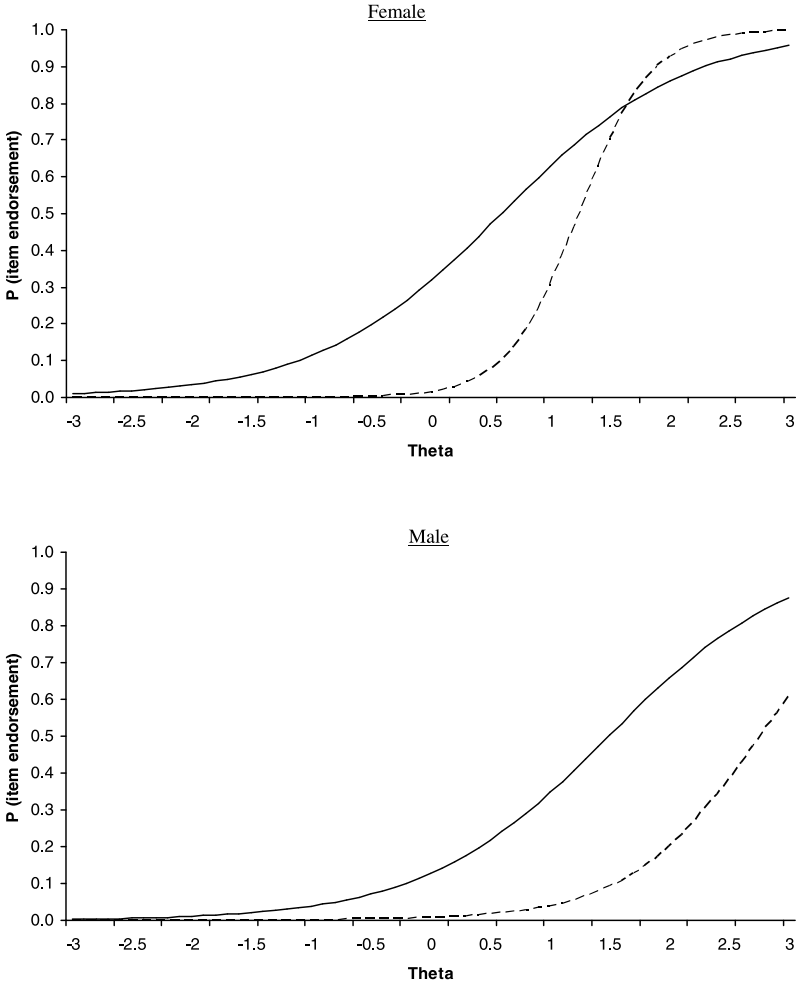

FIGURE 1 Trace lines for "Cries a lot" by age and gender. Solid lines represent trace lines for younger participants (age 2–11); dashed lines represent trace lines for older participants (age 12–17).

Once the discrimination and severity parameters have been estimated for each item in a scale, it is then possible to use these item parameters to compute an estimate of the underlying ability score for each participant (see Thissen & Orlando, 2001). In longitudinal applications where the same scale has been administered to participants on multiple occasions, scores may be calculated for each repeated measure of each participant (i.e., person- and time-specific scores). These IRT scale scores may subsequently serve as the dependent variables in standard longitudinal analyses, such as latent curve analysis (e.g., Bollen & Curran, 2006).

Potential Effects of DIF

For longitudinal modeling results to be valid, it is critical that the item parameters be sufficiently invariant across time. Or, as we elaborate later, if the item parameters show a nontrivial lack of invariance, it is of key importance to adjust the scoring procedure to account for the measurement difference due to time. Furthermore, it may be important to assess measurement equivalence across other relevant study variables, such as gender, or, in the case of cross-study analysis, study membership.[1] Failure to find and account for measurement nonequivalence compromises internal validity and could have undesired consequences for the interpretation of models estimated from scale scores (regardless of whether the scores are calculated based on IRT methods or more traditional methods; e.g., Reise et al., 1993). Because latent curve models represent both latent mean and covariance structures, measurement nonequivalence affecting either observed means or observed covariances can produce misleading results. Thus, accounting for measurement nonequivalence is even more important for these longitudinal analyses. By introducing biases in observed means or covariances, critical substantive results due to developmental status or other covariates could be masked by measurement differences, or conversely, observed developmental or covariate differences might be artifacts of measurement differences rather than true differences.

The nature and size of the impact of measurement nonequivalence on final model interpretations will depend on the direction and effect sizes associated with each item showing DIF and how those DIF effects accumulate across items to affect the scale as a whole. Several researchers have emphasized that statistically significant group bias at the item level does not necessarily translate into practically significant differences in test scores (e.g., Drasgow,

[1]In any given study, there may be many potential covariates that could be subjected to measurement invariance testing. However, we recommend that researchers carefully consider theory and prior psychometric analyses to choose a small number of important covariates for invariance testing.

1987; Roznowski & Reith, 1999; Stark, Chernyshenko, & Drasgow, 2004). For instance, if two separate items have DIF across two groups but in opposite directions (e.g., one item may have stronger discrimination for males relative to females, whereas another may have stronger discrimination for females than males), these DIF effects will cancel each other out to some extent in the creation of scale scores. Additionally, because items with greater discrimination are given more weight in the scoring process than items with lower discrimination, we would expect DIF effects from items with relatively low discrimination values to have little influence on overall scale scores. However, the possible effects of DIF on the scoring of a particular scale and subsequent data analyses using those scores remains a question that should be answered empirically.

On finding significant DIF for a particular item, a researcher may choose to ignore it, drop the offending item from the scale, or adapt a scoring procedure that accounts for the measurement noninvariance across groups. As discussed earlier, depending on the impact of DIF on overall scale scores, ignoring DIF may severely confound measurement differences with true differences (or a lack thereof) in scale scores due to theoretically important covariates such as age. Dropping an item (or more) with DIF from the scale will have the undesirable consequences of altering the scale's content validity and reducing its reliability. Therefore, we recommend the careful assessment of DIF effects, and if these appear nontrivial, we recommend using scoring procedures that explicitly account for these differences in measurement.

Following a brief description of the longitudinal data sources for our cross-study analyses, we present an evaluation of measurement invariance in the CBCL internalizing scale as a function of age, gender, and study using DIF tests for the 2PL model. We then describe a straightforward approach to item parameter estimation and scoring that takes into account the results from the DIF tests, thus maximizing the validity of the IRT scale scores across age, gender, and study for use in subsequent latent curve analysis. Ultimately, we present latent curve model results using scores that do and do not take DIF into account to help demonstrate whether and how DIF is likely to affect such an analysis. Readers can obtain computer code for the analyses by contacting the first author via electronic mail or at www.yorku.ca/dflora.

DATA AND SUBSTANTIVE BACKGROUND

We used data drawn from two existing longitudinal data sets, both coming from studies that use an accelerated longitudinal design to compare the developmental trajectories of various outcomes for children of alcoholics (COAs) with trajectories for children of nonalcoholics (non-COAs). The Michigan Longitudinal Study (MLS) consists of 583 children assessed from one to four times between

ages 2 and 15 (see Zucker & Fitzgerald, 1991; Zucker et al., 2000, for details). The second study, the Adolescent and Family Development Project (AFDP), consists of 443 adolescents assessed from one to three times between ages 10 and 17 (see Chassin, Barrera, Bech, & Kossak-Fuller, 1992; Chassin, Rogosch, & Barrera, 1991, for details). Taken together, the two data sets span a period from 2 to 17 years of age. In both studies, mothers reported on participant internalizing symptomatology using the CBCL. For the current analyses, we used 13 items from the anxiety-depression subscale.[2]

Our ultimate goal for the current analyses is to construct a model for internalizing that spans the full age range covered by the two studies and to test whether and how a model-implied trajectory for COAs differs from that for non-COAs (see Hussong et al., 2008, for theoretical background). Before doing so, however, we establish measurement equivalence across the two studies and across developmental period. We also test for measurement equivalence by gender because prior studies suggest that internalizing-related items often show different measurement properties according to gender (e.g., Schaeffer, 1988).

IRT ANALYTIC METHOD AND RESULTS

Our IRT analyses proceeded according to four stages: dimensionality assessment, item calibration (i.e., IRT parameter estimation) and subsequent scale scoring assuming no DIF across all items, DIF testing, and item calibration and scale scoring accounting for DIF. We thus arrived at two sets of scale scores, one under the assumption of no DIF and one accounting for DIF, for the purposes of our investigation into the effects of measurement nonequivalence.

To conduct DIF testing and item calibration, it was necessary to create a "calibration sample" consisting of independent observations drawn from the repeated measures of the 1,026 participants. In doing so, we sought to keep the calibration sample as large as possible to obtain accurate item parameter estimates while maintaining age heterogeneity to facilitate testing for DIF according to age. We thus created the calibration sample by randomly selecting one observation from each participant's set of repeated observations. (Although a given participant was observed at as many as four different ages, that participant's set of item responses from only one age was selected for item calibration.) The calibration sample consisted of one set of 13 item responses for each of the 1,026 participants.

[2]For both studies, the Likert-type response scale ranged from 0 to 2; however, because of sparse response frequencies at the highest value, we dichotomized each item to represent either the absence (item score = 0) or presence (score = 1) of a given internalizing symptom.

Dimensionality Assessment

As mentioned earlier, an important assumption for the 2PL IRT model is that the set of items is unidimensional. Therefore, prior to conducting our IRT analyses, we conducted exploratory factor analysis (EFA) for the 13 internalizing items, applying a robust weighted least squares estimator to the interitem tetrachoric correlations (see Flora & Curran, 2004). A scree test suggested that there was a single, dominant factor. The root mean squared error of approximation (RMSEA) statistic for the one-factor model was .053, further suggesting reasonable fit of a one-factor model (see Browne & Cudeck, 1993). We obtained similar results when we conducted separate EFAs within each group created for DIF testing.

Item Calibration and Scoring Without DIF

Item calibration is the phase of the analysis where the discrimination and severity parameters of the IRT model are estimated for each item. Because we ultimately seek to analyze data from two studies simultaneously in a single latent curve analysis of CBCL internalizing scores, it is essential that we first establish a common metric for scoring the 13 items from the two studies. A variety of methods, known as common-items equating, are available for scale equating across two data sources using IRT when there is a set of items that is shared by, or common to, the two samples being combined (see Kolen & Brennan, 2004, for a detailed overview).

Here, because the two studies included the same set of 13 internalizing items and because we are initially assuming that there is no DIF due to data source, defining a common scale for the two studies was straightforward. Specifically, to estimate a set of item parameters (and hence establish the internalizing scale) that was constant across the two studies, we simply concatenated the item data from the two samples into a single file, without inclusion of a variable for group membership (i.e., data source). Thus, we treated the data from two samples as if they came from a single source. Technically, this approach is a type of common-items equating procedure called *concurrent calibration* (Wingersky & Lord, 1984).[3] Methodological research has shown that concurrent calibration and other equating methods tend to produce very similar results (Kim & Cohen, 1988). Thus, for this initial item calibration, we assumed that there were no measurement differences across the two studies and also ignored potential DIF by age and gender. We then estimated the 2PL item parameters for each of the 13 items using the full calibration sample of 1,026 with MULTILOG

[3]Unlike other common-items equating methods that rely on a multiple-group formulation, concurrent calibration produces single-population item parameter estimates without the need to transform item parameters according to the mean and variance shift across groups.

TABLE 1
Item Parameter Estimates and Standard Errors Assuming No Differential Item Functioning

		Discrimination		*Severity*	
Item	*P+*	*Estimate*	*SE*	*Estimate*	*SE*
1. Complains of loneliness	.20	1.34	.15	1.38	.13
2. Cries a lot	.20	0.81	.12	1.89	.26
3. Fears might do bad	.17	1.40	.16	1.48	.14
4. Has to be perfect	.43	1.13	.12	0.32	.09
5. Complains no one loves him/her	.24	1.54	.15	1.06	.10
6. Feels worthless	.21	2.36	.23	1.00	.07
7. Nervous/high strung/tense	.21	1.32	.15	1.33	.13
8. Too fearful/anxious	.17	1.89	.20	1.27	.10
9. Feels too guilty	.11	2.58	.30	1.46	.09
10. Self-conscious	.52	1.35	.13	−0.10	.07
11. Unhappy/sad/depressed	.21	2.46	.24	0.97	.06
12. Worries	.37	2.07	.19	0.44	.06
13. Feels others out to get him/her	.10	1.53	.20	1.97	.18

Note. $N = 1{,}026$. $P+$ = proportion of participants with valid data endorsing the item.

(Thissen, Chen, & Bock, 2003). The estimated item parameters are given in Table 1.

In the scoring phase, we used the item parameter estimates from the calibration phase to calculate IRT-scaled internalizing scores for each participant's set of repeated observations based on her or his item responses. In other words, each participant contributed data from only one age (i.e., one randomly selected repeated observation) in the item calibration phase, but then the set of item parameters estimated from the calibration phase was used to calculate scale scores for all ages (i.e., all repeated observations for each participant). Specifically, we estimated maximum a posteriori scale scores (MAPs; see Thissen & Orlando, 2001), as implemented by the scoring function in MULTILOG. In short, a given participant's scale score is estimated from the maximum of the posterior function derived from the product of the trace lines associated with that person's pattern of responses to the 13 internalizing items (along with a standard normal prior density). These MAPs then served as the observed dependent variables in the subsequent latent curve analysis, described later.

DIF testing. Although a variety of methods have been developed for testing DIF (see Wainer, 1993), here, we rely on the likelihood ratio method of Thissen et al. (1993) because this approach has good statistical power and Type I error control (see Wang & Yeh, 2003) and because it easily builds on the basic 2PL model we have discussed so far. For dichotomous items, this method relies on a

multiple-group generalization of the 2PL model, giving the probability of item endorsement for item j in group g as

$$P(y_{jg} = 1|\theta) = \frac{1}{1 + \exp[-1.7a_{jg}(\theta - b_{jg})]}. \quad (2)$$

The location and scale of θ is identified by fixing its mean to 0 and variance to 1 for one of the groups; this group is often called the *reference group,* and a second group for which the mean and variance must be estimated is often called the *focal group.*[4]

DIF testing proceeds by comparing the fit of a model with the discrimination and severity parameters allowed to be free across groups with the fit of a model where these parameters are constrained to be equal across groups. Specifically, the familiar likelihood ratio statistic is calculated from these two models, such that

$$G^2(d.f.) = -2(ll_{\text{Model 1}} - ll_{\text{Model 2}}), \quad (3)$$

where $ll_{\text{Model 1}}$ is the log-likelihood value of the model where item parameters are equal across groups and $ll_{\text{Model 2}}$ is the log-likelihood value of the model where item parameters differ across groups. To test for DIF in a single item j, the value of G^2 is evaluated against a χ^2 distribution with degrees of freedom equal to two if both the discrimination and severity parameter are freed across groups for DIF testing. If it may be assumed that the discrimination parameter does not vary across groups, then it is possible to test for DIF in the severity parameter only, in which case G^2 is evaluated against a χ^2 distribution with 1 *df*. G^2 tests the null hypothesis that the parameters of an item's trace line do not differ between groups; if the statistic is significant, there is sample evidence that the item has DIF.

Often, researchers will choose a subset of items to serve as an "anchor" for which it is assumed a priori from theory or previous research that there is no DIF. The anchor items are constrained to have equal parameters across all groups and thus provide a basis for estimating the group-mean difference on the latent construct (see Thissen et al., 1993). Alternatively, one can test for DIF in each item separately without designating an anchor by fitting a series of models in which a single item is tested for DIF with all remaining items serving as an anchor. We followed the latter procedure for the current analyses because we did not have solid theoretical expectations about which of the 13 items in

[4]Note that it is possible to estimate separate item parameters for several groups simultaneously. However, in this article, our DIF analyses compare only two groups because of sample size considerations.

the CBCL anxiety-depression subscale would display DIF according to age. Using standard IRT software such as MULTILOG, this approach can be quite time consuming given the number of separate models that must be estimated. Fortunately, the freely available software IRTLRDIF (Thissen, 2001) automates the process.

Another potential limitation of testing for DIF across multiple items is that a large number of significance tests accumulates, thus raising concerns about family-wise Type I error. Williams, Jones, and Tukey (1999) discussed the benefits of a procedure by Benjamini and Hochberg (1995) for controlling the false discovery rate associated with multiple tests relative to the well-known Bonferroni multiple comparison adjustment. The Benjamini-Hochberg procedure is easy to implement (see Thissen, Steinberg, & Kuang, 2002) and has been successfully applied in DIF testing contexts (e.g., Steinberg, 2001). In the current analyses we have also used the Benjamini–Hochberg criteria for determining that a given item has significant DIF in the context of DIF testing across all 13 items.

We first tested for age-related DIF in the 13 CBCL items, pooling data across the two studies. Given the goal of assessing measurement invariance as a function of age, ideally we would have tested for DIF across each pair of adjacent ages (i.e., age 2 vs. age 3, age 3 vs. age 4, etc.). This approach is common in educational settings where large samples of participants are commonly sampled within each school grade (e.g., Thissen, Sathy, Flora, Edwards, & Vevea, 2001). However, doing so would have severely depleted our within-group sample sizes given that the full calibration sample of 1,026 consisted of observations drawn at each age from 2 to 17. Thus, we dichotomized the calibration sample into two groups, young (age 2–11; $n = 475$) and old (age 12–17; $n = 551$). In addition to providing ample sample size for each of the two groups, it is important to note that there is also a theoretical basis for this age cutoff based on developmental patterns of internalizing symptomatology (see Angold & Costello, 2001).

Seven of the 13 items showed significant DIF across these age groups. Four items had DIF in both the discrimination and severity parameters: Item 2, "cries a lot"; Item 3, "fears he or she might do something bad"; Item 4, "has to be perfect"; and Item 5, "complains no one loves him or her." With the exception of Item 4, the discrimination parameter was greater for older than younger participants for each of these items, indicating that there is a stronger relation between each item and the underlying internalizing construct among older participants than among younger participants (or, equivalently, the items have a greater amount of measurement error with younger participants). The severity parameter of Items 2 and 5 was greater for older participants, indicating that older participants endorsing these items tended to have higher levels of internalizing than younger participants. For Items 3 and 4, the severity parameter was greater for younger participants, indicating that younger participants endorsing these items tended to have higher levels of internalizing than older participants. Three

items had DIF in the severity parameter only: Item 1, "complains of loneliness"; Item 8, "too fearful or anxious"; and Item 9, "feels too guilty." With Items 1 and 8, the severity parameter was greater, and thus indicative of higher levels of internalizing, for older participants, whereas the severity parameter of Item 9 was greater for younger participants.

Although this DIF testing stage of our analyses produced specific item parameter estimates, these estimates pertain to the situation where parameters are free to vary due to DIF only one item at a time. Therefore, we present specific parameter estimates and discuss effect sizes for these significant DIF tests later in the section on item calibration, where we account for DIF in multiple items simultaneously.

Because we wanted to account for age DIF while testing gender DIF, we implemented a procedure to allow each item with age DIF to have two sets of item parameters, one set for younger participants and one for older participants. Specifically, if item j was characterized by age DIF, we created two new item response variables: Item j-young consisted of responses to item j for younger participants but was set as missing for older participants, and item j-old contained responses for older participants but was missing for younger participants (see Wainer, 1993, p. 130). These new items, which we refer to as *subitems,* then replaced the original item in subsequent analyses, thus accounting for age DIF.

Several items and subitems displayed gender DIF. In particular, the subitem created for young participants from Item 2 ("cries a lot") had a significant gender difference in the severity parameter, which was greater for males than females, but not in the discrimination parameter. The subitem for old participants created from Item 2 had significant gender differences in both discrimination (such that it was greater for females) and severity (such that it was greater for males). Finally, there was also a significant gender difference in the severity of the subitem created for old participants from Item 8 ("too fearful or anxious") such that the severity was greater for females.

As earlier, we created additional subitems to account for gender DIF so that we could next test for DIF according to study membership while accounting for both age and gender DIF. No items or subitems displayed significant study DIF, indicating that any potential sources of measurement nonequivalence across the two data sources were explained by age differences or gender differences.[5]

[5]Our decision to estimate DIF according to age first, followed by gender and study, may appear somewhat arbitrary. However, given that our primary focus here is on the incorporation of DIF in longitudinal analyses spanning several developmental periods, we felt that establishing measurement equivalence according to age was of primary importance. We chose to examine DIF according to study last because it was our hope that any measurement differences across studies would be accounted for by differences in the age and gender distributions of each study.

Item Calibration and Scoring With DIF

We fit the 2PL IRT model to the internalizing item response data from the calibration sample, again using MULTILOG. To account for age and gender DIF, item parameters were estimated separately for subitems created as a result of DIF testing. If an item had significant DIF in the severity parameter but not the discrimination parameter, for parsimony, the discrimination parameter was constrained to be equal across the two subitems created to account for DIF. The results of this item calibration are shown in Table 2.

Following Steinberg and Thissen (2006), we view direct comparisons between item parameter estimates as "the most straightforward presentation of effect size"

TABLE 2
Item Parameter Estimates and Standard Errors Accounting for Differential Item Functioning

		Discrimination		*Severity*	
Item	*P+*	*Estimate*	*SE*	*Estimate*	*SE*
1. Complains of loneliness (young)[a]	.20	1.48	.12	1.06	.12
1. Complains of loneliness (old)[a]	.19	1.48	.12	1.47	.12
2. Cries a lot (young female)[a]	.36	0.76	.10	0.59	.34
2. Cries a lot (young male)[a]	.24	0.76	.10	1.49	.24
2. Cries a lot (old female)	.23	1.98	.40	1.24	.15
2. Cries a lot (old male)	.08	1.07	.32	2.76	.65
3. Fears might do bad (young)	.17	1.09	.22	1.52	.30
3. Fears might do bad (old)	.17	1.91	.27	1.40	.13
4. Has to be perfect (young)	.32	1.64	.23	0.43	.11
4. Has to be perfect (old)	.52	0.79	.15	0.06	.16
5. Complains no one loves him/her (young)	.23	1.32	.22	0.96	.18
5. Complains no one loves him/her (old)	.24	1.96	.26	1.07	.10
6. Feels worthless	.21	2.44	.24	0.98	.06
7. Nervous/high strung/tense	.21	1.34	.15	1.30	.13
8. Too fearful/anxious (young female)[a]	.09	2.00	.16	1.20	.11
8. Too fearful/anxious (young male)[a]	.16	2.00	.16	1.20	.11
8. Too fearful/anxious (old female)[a]	.19	2.00	.16	1.41	.13
8. Too fearful/anxious (old male)[a]	.21	2.00	.16	1.11	.12
9. Feels too guilty (young)[a]	.05	2.62	.24	1.70	.14
9. Feels too guilty (old)[a]	.17	2.62	.24	1.31	.08
10. Self-conscious	.52	1.37	.14	−0.11	.07
11. Unhappy/sad/depressed	.21	2.49	.25	0.95	.06
12. Worries	.37	2.07	.19	0.43	.06
13. Feels others out to get him/her	.10	1.55	.21	1.93	.18

Note. $N = 1{,}026$. P = proportion of participants with valid data endorsing the item.

[a]These subitems had discrimination parameters constrained to be equal across groups as suggested by the differential item functioning results.

(p. 405) for DIF. Differences in the discrimination parameter across groups reflect differences in the change in log odds of item endorsement per unit change in theta (the latent construct measured by the item), holding the severity parameter constant. Severity parameter differences correspond to differences in the observed rates of item endorsement (in population standard deviation units), holding the discrimination parameter constant (see Steinberg & Thissen, 2006). When both item parameters differ across groups, graphical displays offer the most effective means of demonstrating DIF effect size (Steinberg & Thissen, 2006; also see Orlando & Marshall, 2002). Therefore, trace lines for subitems created as a result of significant DIF in both the discrimination and slope parameters are illustrated in Figures 1 to 4.

The trace lines for Item 2, "cries a lot" (Figure 1) are particularly interesting, in that this item had significant DIF as a function of both age and gender. For both males and females, the steepness of the dashed line relative to the solid line indicates that this item was notably more discriminating among older participants than among younger participants. Among younger participants, this item does not appear to be strongly related to the latent construct. This result is intuitively appealing because some younger children may be prone to crying regardless of their standing on the underlying construct of internalizing, whereas frequent crying should be more strongly indicative of psychological distress among adolescents. Furthermore, for both younger and older participants, the trace lines for boys are shifted to the right relative to those for girls, indicating that this item is more likely to be endorsed by girls. This result is consistent with

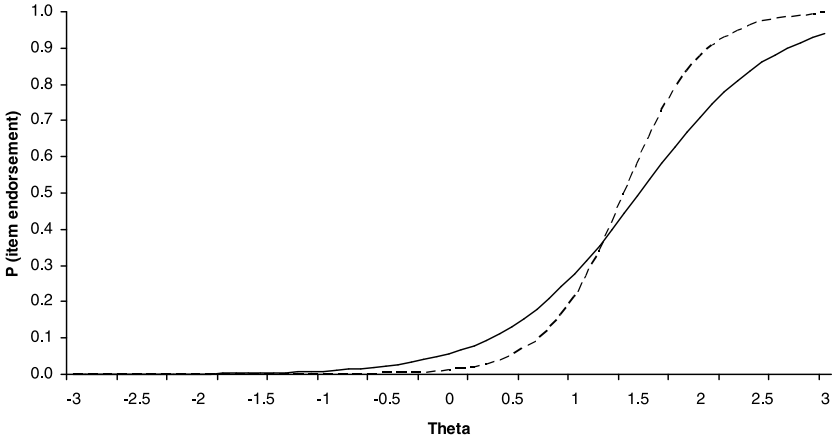

FIGURE 2 Trace lines for "Fears he/she might do something bad" by age. Solid line represents trace line for younger participants (age 2–11); dashed line represents trace line for older participants (age 12–17).

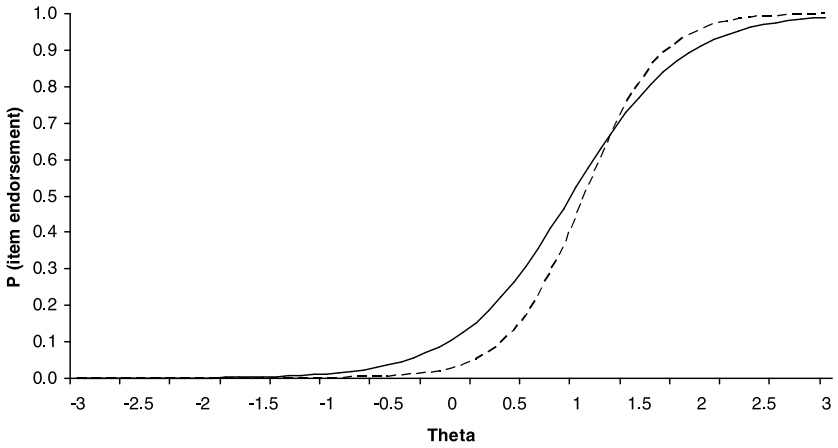

FIGURE 3 Trace lines for "Complains no one loves him/her" by age. Solid line represents trace line for younger participants (age 2–11); dashed line represents trace line for older participants (age 12–17).

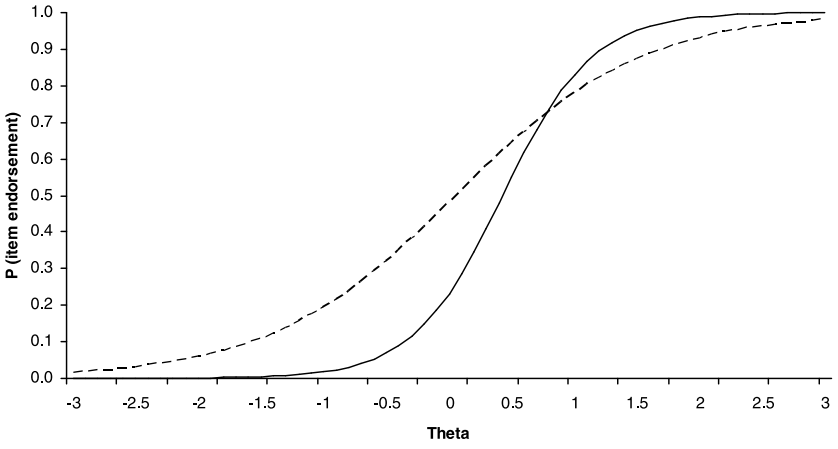

FIGURE 4 Trace lines for "Has to be perfect" by age. Solid line represents trace line for younger participants (age 2–11); dashed line represents trace line for older participants (age 12–17).

previous studies examining gender DIF among items related to crying behavior (e.g., Schaeffer, 1988). The size of the gaps between the solid and dashed trace lines are indicative of the overall DIF effect size for these items; relative to the trace lines in Figures 2 to 4, this item has larger DIF effect size. However, it is important to note that for the male trace lines in Figure 1, although the gap between the solid and dashed lines is notable, both trace lines are at the upper end of the theta continuum (severity = 1.49 for young, 2.76 for old) where there are relatively few participants. Thus, this item primarily serves to distinguish among male participants with particularly high levels of internalizing, especially among older male participants.

The trace lines for Item 3, "fears he/she might do something bad" (Figure 2), and Item 5, "complains no one loves him/her" (Figure 3), are similar in that for lower ranges of the latent internalizing continuum, younger participants are more likely to endorse the item, but for upper ranges of the continuum, older participants are more likely to endorse the item. The trace lines for Item 4, "feels he/she has to be perfect" (Figure 4), show the opposite pattern. Again, the sizes of the gaps between the dashed line and the solid line are indicative of the DIF effect sizes for these items. This DIF effect size is strongest in Figure 4, particularly at the lower end of the latent continuum, whereas in Figures 2 and 3, the separate trace lines appear quite close to each other, indicating small DIF effects.

For items with DIF in only the severity parameter, the calibration results suggest that the DIF effect sizes are quite small. The most notable exception again pertains to the "cries a lot" item. When this item was split into separate subitems for young and old participants, the "young" subitem showed significant gender DIF in the severity parameter but not the discrimination parameter. Thus, the "young" subitem was divided further into two subitems by gender. The severity parameter estimate for the "young female" subitem equaled 0.59, whereas the severity parameter estimate for the "young male" subitem was 1.49. Thus, the gender DIF effect size for these two subitems was such that the rate of item endorsement was nearly 1 *SD* unit lower for boys than for girls (where 1 *SD* unit refers to the assumed population standard deviation of the latent internalizing construct). Other items with significant DIF in only the severity parameter showed DIF effect sizes less than half of 1 *SD* unit (i.e., differences in severity parameters across corresponding subitems of less than .50).

In sum, although we found statistically significant DIF according to age, gender, or both, for 7 of the 13 internalizing items, the effects associated with these measurement differences tended to be small. Thus, this analysis suggests that the general IRT likelihood ratio testing method for DIF detection is associated with strong inferential power. Later, we discuss the implications of these individual, item-level DIF effects for the overall impact of DIF across the scale as a whole.

Next, we calculated scores for the internalizing scale, again as MAPs, that account for these DIF effects. Specifically, these scores were calculated according to the item calibration described earlier, but allowed items with significant DIF to have different parameters according to age or gender (i.e., through the use of subitems). To evaluate the overall impact of DIF on test scores, we then compared these scale scores with those that were created under the assumption of no DIF for any item. Within each age from 2 to 17, these scale scores were highly correlated, which is consistent with other studies that report correlations of IRT-scaled scores that account for DIF with scores that ignore DIF (e.g., Orlando & Marshall, 2002). Despite these high correlations, no comparisons have been made regarding how the use of scores that incorporate DIF might impact model fitting relative to the use of scores ignoring DIF. We explore this issue later, where we report the results of a latent curve analysis using scores that do and do not account for DIF.

LATENT CURVE ANALYSIS

After we estimated IRT-based internalizing scores for each participant's set of repeated measures, we used these scores to estimate a piecewise linear structural equation latent curve model (e.g., Bollen & Curran, 2006, pp. 103–106). We specified an identical model for the internalizing scores that accounted for DIF and for the scores that ignored DIF. Specifically, this model consisted of three latent growth factors: The first linear slope represented latent change in internalizing from age 2 to age 7, the second linear slope represented latent change in internalizing from age 7 to 17, and the intercept, or status, factor represented the level of internalizing at age 12 (see Hussong et al., 2008, for further details). Additionally, these growth factors were regressed on study membership (i.e., whether a given participant came from the MLS sample or the AFDP sample), gender, and parental alcoholism (i.e., COA status, or whether participants came from a family with at least one alcoholic parent).[6] Thus, despite testing for study differences in the measurement phase of our analyses, we still included a covariate for study membership in the growth modeling phase to account for the possibility that participants from the two studies showed different trajectories of internalizing. Additionally, to account for the fact that both studies used an accelerated longitudinal design, leading to a substantial amount of missing data within each age, we estimated the model using the full-

[6]Because all participants contributing data from age 2 to age 7 were from the MLS study, the first slope factor was not regressed on study membership. Here, we report results using only a subset of the covariates considered in Hussong et al. (2008).

information maximum likelihood method (e.g., Arbuckle, 1996) implemented with M*plus* software (L. K. Muthén & Muthén, 1998).

Table 3 presents parameter estimates, standard errors, and associated *p* values for the latent curve models estimated using the IRT scores under the assumption of no DIF and using the IRT scores that account for DIF. Figure 5 presents the model-predicted internalizing trajectories for COAs and controls estimated from the IRT scores assuming no DIF, superimposed over the observed means of those scores. Many, but not all, of the general inferential conclusions are consistent across the two sets of IRT scores. Specifically, there is a significant average increase in internalizing from age 2 to 7 coupled with a significant average decrease from age 7 to 17, regardless of whether scores account for DIF. Further, both sets of scores detected significant effects of study, gender, and parent alcoholism on the latent intercept factor, such that age 12 internalizing was, on average, greater for AFDP participants than MLS participants, greater for females than males, and greater for COAs than for non-COAs. Additionally, both sets of scores detected significant effects of gender and parent alcoholism

TABLE 3
Results of Piecewise Growth Model Fitted to Item Response Theory (IRT) Scores Calculated Under the Assumption of No Differential Item Functioning (DIF) and to IRT Scores Accounting for DIF

	Assuming No DIF			*Accounting for DIF*		
Parameter	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*
Latent intercept factor						
Intercept	.335	.053	<.0001	.337	.052	<.0001
Study effect	−.457	.058	<.0001	−.445	.057	<.0001
Gender effect	−.089	.049	.0679	−.086	.047	.0709
Parent alcoholism effect	.167	.054	.0021	.165	.053	.0020
Residual variance	.402	.025	<.0001	.386	.024	<.0001
First latent slope factor (age 2–7)						
Intercept	.101	.036	.0048	.107	.035	.0025
Gender effect	.043	.031	.1745	.039	.031	.2128
Parent alcoholism effect	−.010	.028	.7151	−.011	.027	.6862
Residual variance	.000[a]			.000[a]		
Second latent slope factor (age 7–17)						
Intercept	−.041	.016	.0131	−.038	.016	.0178
Study effect	.025	.017	.1357	.031	.017	.0601
Gender effect	−.041	.012	.0010	−.042	.012	.0007
Parent alcoholism effect	.022	.013	.0889	.023	.013	.0774
Residual variance	.002	.001	.0320	.002	.001	.0286

Note. $N = 1{,}026$.
[a]This parameter estimate was constrained to zero to arrive at proper solution.

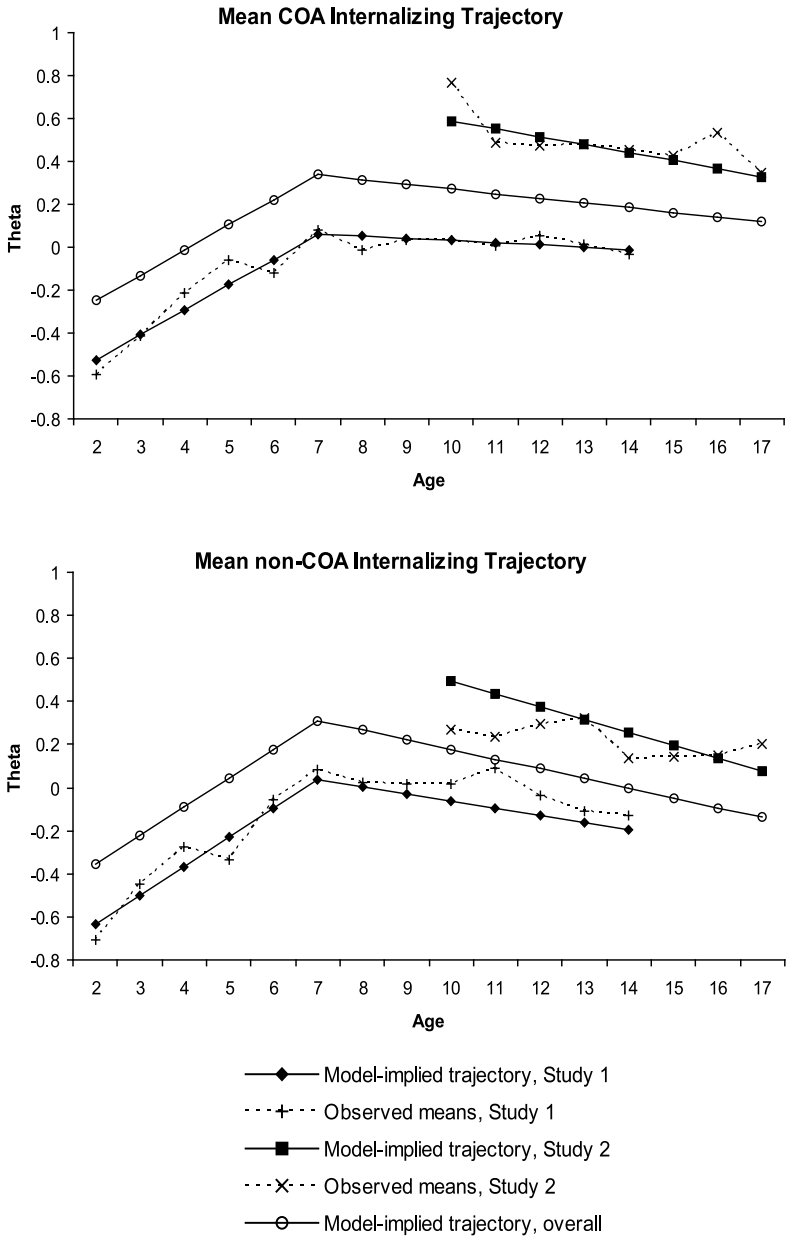

FIGURE 5 Model-predicted trajectories for children of alcoholics and control participants.

on the second slope factor, such that males showed a greater average decrease in internalizing from age 7 to 17 than females and non-COAs showed a greater average decrease than COAs. See Hussong et al. (2008) for further details and conceptual discussion of these results.

The fact that we obtained similar growth modeling results when DIF was ignored relative to when DIF was included is not surprising given that the DIF effect sizes already described were generally small. The only inferential conclusion that differed according to whether the scale scores accounted for DIF pertained to the regression of the second slope factor on the study membership variable. Specifically, the change in internalizing from age 7 to 17 did not significantly vary according to study when the model was estimated using scores that ignored DIF ($p = .14$). Yet, study was marginally significantly associated with this slope factor when DIF was accounted for ($p = .06$), such that internalizing scores decreased more, on average, for participants from the AFDP study than for those from the MFS study.

DISCUSSION

Earlier, we demonstrated a methodology based on IRT that accounts for potential measurement nonequivalence due to age, gender, and data source, building toward a latent curve model fitted to repeated measures of internalizing scores spanning several developmental periods, from early childhood to late adolescence. Our analyses began with formal tests of DIF, followed by the item calibration phase of the analysis, where the 2PL item parameters for the internalizing items were estimated. Importantly, we accounted for measurement nonequivalence by allowing items showing significant DIF to have more than one set of item parameters across the groups for which the item had differing measurement properties (i.e., through the use of subitems). Additionally, we defined a common internalizing scale for the two data sources using concurrent calibration.

Once the item parameters were estimated, we discussed effect sizes describing the extent to which statistically significant DIF was associated with observed differences in item response patterns across age and gender groups. Although several items showed statistically significant DIF, most of the associated effects appeared small. We next estimated IRT scale scores for the repeated measures of internalizing according to the estimates from the item calibration phase. We created two sets of scores: one that accounted for age and gender DIF and another that ignored DIF. Finally, we fitted an identical piecewise linear latent curve model to each set of scores.

In accordance with other researchers (e.g., Khoo et al., 2006), we have stressed the importance of establishing measurement invariance across time in the context of longitudinal studies. However, although there is a relatively

large literature on methods for testing measurement invariance, very few authors have described methods for actually dealing with measurement nonequivalence (i.e., DIF) once it has been found. Thus, the key aspect of our work here is that we have demonstrated a method using IRT for incorporating measurement nonequivalence in the creation of a scale that can be subsequently used in a latent curve analysis or other SEM applications.

When Does Nonequivalence Matter?

We noted that although 7 of the 13 internalizing items showed significant nonequivalence, or DIF, the effect sizes associated with these were small. Accordingly, the particular growth modeling results we presented earlier suggested that whether DIF was accounted for in the creation of internalizing scores had little impact on the inferences drawn from the final growth models that were fitted to these scores. This finding is in line with the assertions of several authors (e.g., Roznowski & Reith, 1999; Stark et al., 2004) that nonequivalence across groups at the item level often does not translate to invariance at the level of the scale as a whole. For instance, earlier we described how the age DIF effect for Item 4, "feels he/she has to be perfect," was in the opposite direction and somewhat larger relative to the DIF effects for Item 3, "fears he/she might do something bad," and Item 5, "complains no one loves him/her." Thus, when scores are calculated for a scale incorporating both of these items, the DIF effects cancel out to some extent. Additionally, we already argued that only one of the items, "cries a lot," showed substantial DIF effect sizes. As this is but one of 13 items contributing to the scale, we might expect that its DIF effect, however large, might have relatively little influence on overall scale scores. Finally, because items with greater discrimination are given more weight in the scoring process than items with lower discrimination, we would expect DIF effects from items with relatively low discrimination values to have little influence on overall scale scores. Here, most of the items with large discrimination values (e.g., > 2.00) did not have statistically significant DIF. Therefore, for these reasons, we did not expect that accounting for DIF would lead to substantially different scores for the internalizing scale, despite the fact that 7 of the 13 items had statistically significant DIF.

Conversely, measurement nonequivalence will affect overall scale scores when a large proportion of items have large DIF effects (i.e., large differences in discrimination or severity parameters across groups) that occur in the same direction for each item. For example, in the preceding analyses, if we had found that severity parameters were consistently lower, by a relatively large amount, for younger participants relative to older participants across all items showing DIF, then we would expect this source of nonequivalence to have a substantial effect on the scale as a whole. As a result, ignoring these DIF effects would then distort the latent curve model representation of true change

in overall internalizing behavior across developmental periods. Or, if we had found that substantial DIF in discrimination parameters such that several items were consistently more discriminating among females than among males, this effect would have ramifications for the reliability of the internalizing scale because of the relationship between item discrimination and measurement error. Ignoring this DIF effect would then affect the statistical power for finding gender differences in the latent curve model. Nonetheless, even subtle DIF effects on scale scores can affect parameter estimates, potentially leading to different inferential conclusions. For instance, accounting for small DIF could determine whether a particular p value reaches statistical significance or just misses significance (e.g., $p = .051$ vs. .049). Additional work is needed to delineate more clearly the situations under which measurement nonequivalence due to age or other factors is most likely to influence results of latent curve analyses or other SEM applications.

Another important aspect of our analyses is that we have combined data from two studies to form what may be called a "cross-study" analysis. Just as we emphasized the establishment of a common scale as a function of age, we also noted the necessity of establishing a common scale across these two data sources. As mentioned already, we did not find that any of the internalizing items had DIF across studies. However, if an item had DIF according to data source, we would have incorporated this DIF by estimating separate item parameters for the two data sources, just as we did for items showing significant age or gender DIF. It was also relatively simple for us to place the data from the two studies onto a common scale because the same 13 internalizing items were administered in both studies. Yet, in other applications, it may be that the studies being combined administer different item sets for the measurement of the same theoretical construct. In this case, it is still theoretically possible to define a common scale across studies as long as there remain items in common (see Kolen & Brennan, 2004).

There are potential limitations to our approach. For instance, we tested for age-related DIF among the internalizing items by dichotomizing our calibration sample into only two groups, young and old, potentially removing more subtle effects that may occur across shorter developmental periods. In that DIF testing methods rely on multiple-group item response models, it was necessary to treat age as a discrete variable. However, as already mentioned, an ideal analysis would have tested for DIF across more age groups. Still, we felt that our available sample sizes would not accurately support a more fine-tuned assessment of DIF as a function of age. Similarly, it is theoretically possible to test for age and gender differences simultaneously by creating four groups from crossing the young and old groups with the male and female groups, and then estimating four-group rather than two-group IRT models. However, again because of sample size restrictions, we preferred to test age and gender DIF separately.

Another potential limitation to our analyses is that we have proceeded using a "two-stage" modeling process for our repeated measures of internalizing symptomatology. That is, the first stage of our process consisted of the creation of an internalizing scale and estimation of scale scores using methods from IRT. These scores were then saved and used as the observed dependent measure for the second stage of the modeling process, which was the fitting of latent curve models. Thus, the second stage of the analyses implicitly assumed that the scale scores from the first stage were calculated without error. Nonetheless, each of the IRT scale scores has an associated standard error (see Thissen & Orlando, 2001), which ideally would be included in the growth model. Other researchers have suggested approaches for simultaneously including a measurement model for the individual items within a higher order growth model for the scale as a whole (e.g., Bollen & Curran, 2006, pp. 247–251). However, these methods are currently difficult to implement in practice given the large number of parameters that must be simultaneously estimated.

IRT versus CFA

IRT and CFA share the common goal of modeling observed variables as a function of a latent construct.[7] Yet, when the observed variables for a CFA are dichotomous, as item responses often are, the usual approach of estimating the model from product–moment correlations or covariances leads to inaccurate findings (West, Finch, & Curran, 1995). Instead, one method that explicitly accounts for the categorical nature of the variables estimates the CFA model from the set of univariate thresholds and tetrachoric correlations calculated from the marginal and joint distributions of the observed dichotomies (B. Muthén, 1984; Flora & Curran, 2004). Takane and de Leeuw (1987) showed that a single-factor CFA model estimated in this fashion leads to parameters that are mathematically equivalent to the two-parameter normal ogive (2PNO) IRT model.[8] That is, following formulas provided by Takane and de Leeuw, there are one-to-one relationships between the IRT discrimination parameter for a given item and its CFA factor loading and between its IRT severity parameter and its threshold parameter from the dichotomous CFA model.

Just as there is an extensive literature on DIF in IRT, there are also many resources on invariance testing with CFA. In CFA, invariance testing can be

[7]Just as CFA can be extended to include more than one factor, additional latent variables can also be incorporated into IRT models through the use of multidimensional IRT (e.g., Bock, Gibbons, & Muraki, 1988).

[8]Whereas we are using the logistic form of the two-parameter model, the scaling constant (1.7) can convert the logistic parameters back to the normal ogive metric. When this scaling constant is used, the trace lines generated by these two models are virtually indistinguishable (Thissen & Orlando, 2001).

conducted using either a multiple-group approach (e.g., Millsap & Yun-Tein, 2004) or using the multiple indicator, multiple causes (MIMIC) approach (B. Muthén, 1989). The multiple-group approach proceeds in a manner very similar to that described earlier for IRT in that likelihood ratio tests compare models with item parameters constrained across groups to models with parameters freed across groups. Given the formal relationships between CFA and IRT, very similar findings should be obtained across these two multiple-group methods. In the MIMIC approach, the grouping variable is directly included in the model as a covariate predicting both the latent trait and the individual items. A significant relation between the covariate and an item response, controlling for the latent trait, is indicative of threshold DIF (i.e., in a severity parameter).

Despite these parallels between CFA analyses of measurement invariance and IRT analyses of DIF, here we demonstrate the IRT approach for several reasons. First, by scaling the latent variable according to a standard normal prior, IRT scale scores are placed on a meaningful, well-understood metric that incorporates the severity concept in a relatively clear way. Additionally, IRT utilizes "full-information" estimation methods that simultaneously estimate all model parameters directly from the observed data rather than from summary statistics, whereas the CFA approach previously described relies on "limited-information" estimation where thresholds, tetrachoric correlations, and model parameters are separately estimated in a three-stage procedure, such that inaccuracies at an earlier stage can affect estimates at a later stage (see Wirth & Edwards, 2007). Furthermore, although we have demonstrated the 2PL model here, IRT is more flexible than CFA in terms of the item response formats that can be accommodated. Specifically, the three-parameter logistic IRT model is well suited for items where a correct response can be guessed, whereas the nominal IRT model is appropriate for items with unordered response options (see Embretson & Reise, 2000). Next, a disadvantage of the MIMIC approach in particular is that group differences in factor loadings, or item discrimination, cannot be tested in any straightforward fashion (see Finch, 2005). As already reported, several of the CBCL items had significant DIF in their discrimination parameter. Finally, IRT has historically developed as a method that heavily relies on presenting results graphically, which can be particularly useful for examining DIF (Raju, Laffitte, & Byrne, 2002; Steinberg & Thissen, 2006). We illustrated the use of IRT trace line plots for examining DIF with IRT earlier.

SUMMARY

Throughout this article, we emphasize the importance of considering the possible distortion of results that can occur from a lack of measurement equivalence. This consideration is particularly important in longitudinal studies where mea-

surement properties often change as a function of respondents' developmental period (e.g., Patterson, 1993) or in cross-study analyses where data from two different studies are combined. In many cases, researchers can use well-validated instruments for which there is previous research regarding measurement equivalence across developmental periods. Often, however, whether an instrument's psychometric properties remain invariant as a function of development or other covariates is unknown.

Therefore, we recommend formal evaluation of measurement equivalence using either the IRT-based methods described earlier or using categorical CFA methods. If a statistically significant lack of measurement equivalence is found, it is important to consider whether the associated effect sizes are nontrivial. We described how DIF effect sizes can be evaluated at the item level. A relatively simple way to determine whether these DIF effect sizes are likely to have an impact at the overall scale level is through the use of test characteristic curves (Thissen, Nelson, et al., 2001). These plots can reveal the extent to which the relationships between the underlying construct and the expected sum score on a measure of the construct are likely to vary across two groups for which significant DIF has been found (e.g., Orlando & Marshall, 2002). If the overall effect of DIF across the scale is nontrivial, it becomes critical that scale scoring procedures account for this issue. Failure to do so will compromise the validity of subsequent analyses, such as latent curve modeling, that are based on these scores.

ACKNOWLEDGMENTS

The authors received support from grants DA15398 (Andrea M. Hussong) and DA13148 (Patrick J. Curran); contributing studies were supported by grants AA07065 (Principal Investigators [PIs]: Zucker & Fitzgerald) and AA016213 (PI: Chassin). We thank Dan Bauer, Li Cai, Daniel Serrano, and R. J. Wirth for their valuable input.

REFERENCES

Achenbach, T., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and revised Child Behavior Profile.* Burlington, VT: University Associates in Psychiatry.

Angold, A., & Costello, E. J. (2001). The epidemiology of depression in children and adolescents. In I. M. Goodyear (Ed.), *The depressed child and adolescent* (2nd ed., pp. 143–178). New York: Cambridge University Press.

Arbuckle, J. (1996). Full information estimation in the presence of incomplete data. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57,* 289–300.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* Hoboken, NJ: Wiley.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of testing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Chassin, L., Barrera, M., Jr., Bech, K., & Kossak-Fuller, J. (1992). Recruiting a community sample of adolescent children of alcoholics: A comparison of three subject sources. *Journal of Studies on Alcohol, 53,* 316–319.

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

Curran, P. J., Edwards, M. C., Wirth, R. J., Hussong, A. M, & Chassin, L. (2007). The incorporation of categorical measurement models in the analysis of individual growth. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29,* 278–295.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Hussong, A. M., Flora, D. B., Curran, P. J., Chassin, L., & Zucker, R. (2008). Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents. *Development and Psychopathology, 20,* 165–193.

Khoo, S. T., West, S. G., Wu, W., & Kwok, O.-M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 301–317). Washington, DC: American Psychological Association.

Kim, S., & Cohen, A. S. (1988). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 116–130.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

McArdle, J. J., & Horn, J. L. (2002, October). *The benefits and limitations of mega-analysis with illustrations for the WAIS.* Paper presented at the international meeting of CODATA, Montreal, Canada.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39,* 479–515.

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

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585.

Muthén, L. K., & Muthén, B. O. (1998). *M*plus *user's guide* [Computer software]. Los Angeles: Muthén & Muthén.

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. *Psychological Assessment, 14,* 50–59.

Patterson, G. R. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology, 61,* 911–919.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59,* 248–269.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34,* 100–114.

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. In C. C. Clogg (Ed.), *Sociological methodology* (Vol. 18, pp. 271–307). Washington, DC: American Sociological Association.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16,* 41–49.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item functioning and differential test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89,* 497–508.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81,* 332–342.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11,* 402–415.

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

Thissen, D. (2001). IRTLRDIF v. 2.01b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). MULTILOG (version 7) [Computer software]. Chicago: Scientific Software.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141–186). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Thissen, D., Sathy, V., Flora, D. B., Edwards, M. C., & Vevea, J. L. (2001, January). *Scaling the new EOG mathematics and computer skills tests.* Presentation at the North Carolina Accountability Testing Conference, Greensboro, NC.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27,* 77–83.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123–135). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27,* 479–498.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56–75). Newbury Park, CA: Sage.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24,* 42–69.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79.

Zucker, R. A., & Fitzgerald, H. E. (1991). Early developmental factors and risk for alcohol problems. *Alcohol Health and Research World, 15*(1), 18–24.

Zucker, R. A., Fitzgerald, H. E., Refior, S. K., Puttler, L. I., Pallas, D. M., & Ellis, D. A. (2000). The clinical and social ecology of childhood for children of alcoholics: Description of a study and implications for a differentiated social policy. In H. E. Fitzgerald, B. M. Lester, & B. S. Zuckerman (Eds.), *Children of addiction: Research, health and policy issues* (pp. 109–141). New York: Routledge Falmer.