Published in final edited form as:
Psychol Methods. 2009 June ; 14(2): 101–125. doi:10.1037/a0015583.

Psychometric Approaches for Developing Commensurate Measures Across Independent Studies: Traditional and New Models

Daniel J. Bauer and **Andrea Hussong**
The University of North Carolina at Chapel Hill

Abstract

When conducting an integrative analysis of data obtained from multiple independent studies, a fundamental problem is to establish commensurate measures for the constructs of interest. Fortunately, procedures for evaluating and establishing measurement equivalence across samples are well developed for the linear factor model and commonly used item response theory models. A newly proposed moderated nonlinear factor analysis model generalizes these models and procedures, allowing for items of different scale types (continuous or discrete) and differential item functioning across levels of categorical and/or continuous variables. The potential of this new model to resolve the problem of measurement in integrative data analysis is shown via an empirical example examining changes in alcohol involvement from age 10 to 22 across two longitudinal studies.

Integrative Data Analysis (IDA), or the simultaneous analysis of data obtained from two or more independent studies, offers many potential advantages. As reviewed by Curran & Hussong (this issue), correlated advantages of IDA include economy (i.e., reuse of extant data), power (i.e, large combined sample sizes), the potential to address new questions not answerable by a single study (e.g., combining longitudinal studies to cover a broader swath of the lifespan), and the opportunity to build a more cumulative science (i.e., examining the similarity of effects across studies, and potential reasons for dissimilarities). There are also many methodological challenges associated with IDA, including the need to account for historical effects, regional differences, and sample heterogeneity across studies (Curran & Hussong, this issue). Perhaps the most fundamental challenge, however, is the need to construct commensurate measures across studies for both predictors and outcomes. It is this task that is the central focus of our paper.

Our paper takes the following course. First, we review the nature of the problem of measurement in IDA and comment on ad hoc solutions that have previously been employed for putting measures on common scales across studies. Second, we argue that psychometric tools developed to assess measurement equivalence in other contexts can also be applied to facilitate IDA, but that certain extensions of the underlying measurement models will often be necessary. In particular, we provide a cursory review of the linear factor analysis model and the 2-parameter logistic item response theory model, including the topics of factorial invariance and differential item functioning. We then propose a new, more general measurement model that we believe offers several additional advantages for IDA. Third, we apply this new approach

to evaluate changes in alcohol involvement over adolescence and young adulthood in data pooled from two independently conducted longitudinal studies. Fourth, we discuss the limitations of our approach and areas in need of further methodological development.

The Problem of Measurement

Simply put, without common measures, IDA is a nonstarter. To simultaneously analyze data from multiple studies, measures must be available from each study that reflect the same construct, the construct must have the same theoretical meaning, and the scores for the construct must be scaled commensurately. Clearly, if the same theoretical constructs are not measured in each study, then an IDA involving those constructs cannot be conducted. Constructs given the same label (i.e., "stress"), but measured based on different theoretical and operational definitions (i.e., mental versus somatic, affective versus physiological), are similarly incomparable in IDA. It is when the definition and measurement of constructs agrees across studies that IDA becomes possible.

In an (impossibly) ideal world, each study in an IDA would measure the same constructs using the same, gold-standard instruments. More realistically, however, each study is likely to measure the same constructs differently, particularly if IDA was not planned at the outset. This situation presents a fundamental challenge for IDA: How to put disparate measures of the same construct on a common metric. The problem is, of course, not unique to psychology. Statistical matching, a somewhat different approach to IDA developed within econometrics, also emphasizes the need to *harmonize* measures across data sets (D'Orazio, Di Zio & Scanu, 2006). Harmonization refers to the recoding of variables so that they are scored with identical values in each study. For instance, if one study recorded family income as an open-ended continuous response, whereas a second study used an ordinal scale of <$10,000, $10,000-19,999, $20,000-49,999, etc, then the two measures could be "harmozined" by recoding the continuous response into the ordinal categories. Measurement harmonization is thus a straightforward, face valid attempt to create commensurately scaled variables.

Harmonization may be a useful first step in developing commensurate measures for IDA, but it is not sufficient. There is no guarantee that the harmonized values of a variable are truly equivalent across studies. Although the scoring of the variable may be superficially identical, how these scores map onto levels of the underlying construct may nevertheless differ. Even for the simple example of reporting family income, it is possible that individuals under-report their income when asked to provide a free response, but are less tempted to do so when asked to indicate an income range. Although the continuous measure of income can be recoded to match the ordinal income ranges, identical harmonized values (e.g., $10,000-19,999) may then reflect somewhat higher levels of socioeconomic status in the first study than the second. Similar problems may arise from subtle variations in item prompts or response category labels (Steinberg, 2007; Rivers, Meade & Fuller, in press). Indeed, even identical items which do not require harmonization may function differently across studies, due to differences in regional interpretations, historical periods, or even placement of the items within study-specific test batteries (Holland & Dorans, 2006).

Fortunately, there is a long psychometric literature on evaluating measurement equivalence that can be brought to bear on this problem, provided that there is a set of identical or harmonized items available for each construct across studies. Borrowing the parlance of item response theory, we will refer to these items as *common items*, allowing also for the possibility of non-common items that cannot be reasonably harmonized across studies. Given multiple common items, the psychometric approach will allow us to formally test whether these items function equivalently across studies, permitting IDA to take place. For instance, suppose that a total of 20 items are administered across three studies to assess internalizing behavior, but

only 10 of the items are common items that are shared (or harmonized) across studies. If even some of the 10 common items measure internalizing behavior equivalently across the three studies, then they can be used to ensure that the scores for internalizing behavior are scaled commensurately across studies. The non-common items then provide extra information for estimating the internalizing behavior scores in each study.

Historically, most psychometric models have focused on a single type of item (e.g., continuous or binary), and allowed for tests of measurement inequivalence across two or more independent groups. For IDA, this historical framework will take us far, allowing us to establish commensurate measurement across independent studies so long as the items are of the same type. But the traditional framework can sometimes be too limiting. For instance, in IDA we will often be in the position of cobbling together measures for a construct from items that weren't originally part of the same inventory (e.g., frequency of heavy drinking from a questionnaire and diagnosis of alcohol dependence from a structured interview) and hence may be heterogeneous with respect to type (e.g., a mix of continuous, count, ordinal, and/or binary items). Furthermore, if we conduct longitudinal IDA we may also need to evaluate whether items function equivalently at all ages. Such considerations will take us beyond the traditional framework. We shall thus begin by considering two traditional modeling approaches, and then expand beyond these to accommodate extra complexities that may often arise in IDA applications.

Traditional Psychometric Models

In this section we review two traditional psychometric models that can be used to evaluate measurement equivalence across samples (e.g., studies), the Linear Factor Analysis (LFA) model and the 2-Parameter Logistic (2-PL) Item Response Theory (IRT) model. Both models may be viewed as variants of latent trait analysis (Lazarsfeld & Henry, 1968; Bartholomew & Knott, 1999), differing superficially in whether the items are continuous (LFA) or binary (2-PL). Historically, however, LFA and 2-PL IRT models have been applied for different purposes (e.g., dimensionality testing versus ability scoring) with different data structures (e.g., subtests versus items as indicators) using different estimation approaches (e.g., analysis of summary statistics versus raw response patterns), leading to subtle differences in terminology and practice. Without intending to further reify the somewhat artificial distinction between factor analysis and IRT, we thus review the LFA and 2-PL model separately, followed by a consideration of the potential application of these models to solve the measurement problem in IDA.

The Linear Factor Analysis Model

In the long and broad literature on factor analysis, one topic that has received sustained attention is the issue of measurement equivalence or, as it is often referred to within this context, factorial invariance (see Millsap & Meredith, 2007, for a historical review). As an entry point into this topic, let us consider a LFA model in which a single latent factor is posited to underlie a set of observed, continuous variables (indicators, or items). The model may be written as

$$y_{ij}=\nu_i+\lambda_i\eta_j+\varepsilon_{ij} \quad (1)$$

where i indexes the item (from $i = 1, 2,\ldots, p$) and j indexes the person (from $j = 1, 2,\ldots, N$). Note that Equation 1 is a simple linear regression of each observed item y on the latent factor η, with the intercept and slope (loading) parameters indicated by ν and λ, and the residuals indicated by ε. We will refer to ν and λ jointly as item parameters. Additionally, we will assume that the factor and residuals are normally and independently distributed as $\eta_j \sim N(\alpha, \psi)$ and

$\varepsilon_{ij} \sim N(0, \sigma_i^2)$. [1] The indicators are thus assumed to be *conditionally independent* (or locally independent), given the latent factor. The covariances among the measured items then reflect only the common influence of the latent factor η (i.e., $COV\,(y_{ij}, y_{i'j}) = \lambda_i \lambda_{i'} \psi$ for all $i \neq i'$).

The factor model is said to be identified if a unique solution can be obtained for each parameter. As long as there are at least three indicator variables, the one factor model is identified up to the scaling of the latent factor (Bollen, 1989). The mean and variance of the latent factor are arbitrary (given that it is an unobserved variable), so the scale of the latent variable must be set by the analyst. One common strategy for scaling the factor is to set the intercept and factor loading of a 'reference item' to zero and one, respectively. This identifies the model, places the latent variable on the same scale as the reference item, and allows one to freely estimate the factor mean and variance. Another strategy is to set the mean and variance of the factor to convenient values. For instance, setting the factor mean and variance to 0 and 1, respectively, identifies the model and puts the factor on a standard normal scale.

When the same instrument, or set of items, is administered to two or more independent groups (e.g, male/female, ethnic groups, nationalities) it is often of interest to compare the groups on the latent factor. To validly make such comparisons, one must first ensure that the latent factor is equivalently defined and on a common metric across the groups. In general, it must be possible to hold the item parameters (intercepts and loadings) equal over groups for the factor means and variances to be directly comparable. That is, the observed variables must have the same relationship to the latent factor in each group. When the item parameters are all equal over groups, this is known as *strong factorial invariance* (Meredith, 1993). Often this equality holds only for a subset of items, resulting in *partial invariance* (Byrne, Shavelson & Muthén, 1989; Reise, Widaman & Pugh, 1993).

In practice, factorial invariance is evaluated via the multi-sample factor analysis model developed by Jöreskog (1971) and Sörbom (1974). A LFA model is fit simultaneously in two or more independent samples, and across-group equality constraints on the item parameters are evaluated using Likelihood Ratio Tests (LRTs; see Widaman & Reise, 1997). Briefly, an LRT is computed by taking twice the difference in the log-likelihoods of the two models (with and without equality constraints). This difference is distributed as a chi-square with degrees of freedom equal to the number of equality constraints evaluated. A significant result indicates that the equality constraints should be rejected; otherwise, they are retained. Even if equality constraints are untenable for the full set of items, it may still be possible to retain these constraints for a subset of the items, resulting in partial invariance. Several methods for determining which items are non-invariant have been suggested (see Cheung & Rensvold, 1999). One common strategy is to use Modification Indices (MIs, also known as Lagrange Multiplier tests), which indicate the expected improvement in model fit associated with freeing each constraint in the model (Reise, Widaman & Pugh, 1993; Yoon & Millsap, 2007).

Given the potential for partial invariance, the method used to identify the model can be quite important (Cheung & Rensvold, 1999). Use of the reference item method (i.e., constraining the intercept and loading of one item to 0 and 1 in both groups) can be problematic, as it implicitly assumes that the reference item is invariant across groups. If it is not, tests for invariance for the remaining items will be incorrect. Similarly, one would not want to constrain the mean and variance of the factors to 0 and 1 in both groups, as factor mean and variance differences might then masquerade as intercept and factor loading differences, respectively. Another option is to set the factor mean and variance to 0 and 1 in one group, then estimate

[1] Note that the normality assumption is used to motivate the maximum likelihood estimator for the model and to contrast with subsequent models. Under certain conditions this estimator is consistent even when the residuals and factor(s) are not normally distributed (Browne, 1984).

the factor mean and variance in the other group while placing equality constraints on one or more item intercepts and loadings (Reise, Widaman & Pugh, 1993). Yoon & Millsap (2007) implemented this approach by first placing equality constraints on all items and then releasing these constraints, item by item, based on MIs until no further improvement in fit could be achieved. They noted that this procedure works well when the number of non-invariant items is small relative to the number of invariant items. Thus, as the number of invariant items increases, so too does our confidence that we are measuring the same latent factor on the same scale in each group, permitting across-group comparisons.

The 2-Parameter Logistic Model

As with factor analysis, the development of commensurate measures has long been a focus of item response theory (Holland, 2007). To motivate the 2-PL IRT model we shall suppose that there is a single latent trait or ability that underlies a set of binary item responses (e.g., mathematics proficiency, as reflected by a series of items scored correct or incorrect). Some items may be more difficult than others, and some items may be more strongly related to the latent trait than others. The 2-PL model allows for such differences, assuming a conditional Bernoulli (i.e., binary) distribution for each item in which the probability of observing a score of one (e.g., a correct response) is given by the function

$$P(y_{ij}=1|\theta_j)=\frac{1}{1+\exp\left[-a_i(\theta_j-b_i)\right]} \quad (2)$$

where θ_j is the latent trait score for individual j, and a_i and b_i are referred to as the discrimination and difficulty parameters for item i, respectively, or item parameters, collectively.[2] The item responses are assumed to be independent across individuals and conditionally independent across items given the latent trait. Additionally, it is typically assumed that the latent trait is standard normal, or $\theta_j \sim N(0, 1)$. Note that this convention parallels one of the approaches used in the linear factor analysis model to set the scale of a latent factor. With this assumption in place, the discrimination and difficulty parameters can be estimated for each item.

The parameters of the 2-PL model have appealing interpretations. The difficulty parameter (b) indicates the level of the latent trait at which there is a .5 probability for the item to be scored one, (i.e., $P(y = 1|\theta = b) = .5$). That is, one would expect 50% of examinees with latent trait levels equal to the difficulty value to answer the item correctly. In contrast, the discrimination parameter (a) indicates how rapidly the probability of answering correctly increases as the latent trait increases. The role of these two parameters can also be seen by plotting the Item Characteristic Curves (ICCs, also known as tracelines), or $P(y = 1|\theta)$ as a function of θ, obtained by substituting the known or estimated values of a and b into Equation 2 for each item. Example ICCs for four items are depicted in Figure 1. The values of the difficulty parameters for the four items are indicated by the vertical droplines, showing the level of θ (in standard deviation units) at which each traceline attains a probability value of .5. Note that as difficulty values increase, the ICCs move toward the right of the plot, indicating that higher levels of the latent trait are required for the items to be scored one with appreciable probability. The slopes of the ICCs are then determined by the values of the discrimination parameters. Items with higher discrimination have steeper slopes, whereas those with lower discrimination have flatter slopes. The items with steeper slopes are more highly related to the latent trait.

[2]There is also a direct correspondence between the 2-PL model and the factor analysis model for binary indicators (Kamata & Bauer, 2008; Takane & DeLeeuw, 1987; Wirth & Edwards, 2007). Millsap and Yun-Tein (2004) provide a discussion of invariance testing with binary items from a factor analytic perspective.

Scoring is a major emphasis of IRT, and often the primary motivation behind applying a 2-PL model. Although a variety of scoring methods have been developed, we shall focus on the commonly used Expected a Posteriori (EAP) and Maximum a Posteriori (MAP) scoring methods (see Thissen & Orlando, 2001). An EAP is the expected latent ability of person *j* taking into account his or her correct and incorrect responses and the overall ability distribution in the population. More technically, the EAP score is defined as the mean of the posterior ability distribution for person *j* given *j*'s vector of item responses. Similarly, a MAP is the mode of the posterior ability distribution for person *j*. Both EAPs and MAPs are "shrunken" estimates, meaning that the score obtained for person *j* will be closer to the marginal mean of the overall ability distribution (across persons) as the amount of information available for person *j* decreases. For instance, if a participant had missing values for all items, then the best score estimate for this person would be the overall average, whereas with more non-missing values a more individualized estimate could be obtained. In general, the inclusion of more items, and more informative items, results in less shrinkage and greater certainty about an individual's ability level.

Although the 2-PL model is often described within the context of ability testing, the same model is of course relevant for the measurement of other psychological constructs, such as dimensions of personality or psychopathology. Considering the latter case, the latent trait might represent liability to pathology, measured by a set of binary symptom indicators (i.e., symptom present or absent). Higher "difficulty" values would indicate higher symptom severity (i.e., that a higher level of underlying psychopathology is necessary to observe the symptom).

Within the context of ability testing, in particular, a great deal of attention has been paid to establishing the equivalence of different tests and test forms. Indeed, psychometricians have developed many designs and procedures for linking and equating test scores (see Holland & Dorans, 2006). This emphasis is understandable: if a high stakes testing program uses multiple test forms to increase security, the test scores obtained from each form should have an equivalent meaning and metric. Similarly, if one wishes to track growth in an ability, then the test scores obtained at different ages should be on the same scale, despite the necessity of changing the items administered from one age to the next to make the test age-appropriate. Within this context, one design of particular relevance for our purposes is the common-item (or anchor-test) design, in which groups of examinees are given different tests (or test forms) which contain a subset of common items. Similar scenarios are likely to arise frequently in IDA.

Given the presence of a set of common items, one can directly test measurement equivalence by evaluating whether individuals from different groups with equal values for the latent trait have an equal probability of scoring the item correct. When this probability differs over groups for a given item, this is referred to as Differential Item Functioning (DIF). A variety of methods are available to evaluate DIF (see Angoff, 1993), many of which are primarily applicable when there are a large number of examinees and a large number of items (e.g., the standardization method or the Mantel-Haensel statistic). Here, we focus only on the direct IRT method in which the parameters defining the ICCs (e.g., the *a* and *b* parameters in the 2-PL model) are directly contrasted across groups by fitting a multi-sample IRT model (Thissen, Steinberg & Wainer, 1993). Our reasons for focusing on the IRT method are that it is both theoretically elegant and more practically useful when the number of items and examinees is smaller than is typically available within large-scale testing programs (Bock, 1993).

As described by Embretson & Reise (2000, p. 252-262), in the IRT method, the latent trait mean and variance are set to 0 and 1 in the reference group, and are freely estimated in the focal group. The mean and variance of the latent trait in the focal group are identified so long as a subset of the common item parameters are held equal across groups. To determine DIF,

one strategy is to initially constrain all item parameters to equality across groups. Then, one item at a time, the *a* and *b* parameters can be allowed to differ by group. LRTs may be conducted to determine whether allowing for different *a* and *b* parameters significantly improves the fit of the model. A significant LRT is indicative of DIF. Note that this strategy is quite similar to the one recommended for evaluating partial invariance in the linear factor model by Reise, Widaman & Pugh (1993) and Yoon & Millsap (2007). Further heightening the similarity, Glas (1998) suggested using MIs (rather than LRTs) to detect DIF in multi-sample IRT models.

Potential Application In IDA

The procedures reviewed above for the LFA and 2-PL models can be directly applied within the context of IDA to place scores on the same measurement scale across studies. Suppose that all items for a factor (or trait) are common items, available in each study. Then a muti-sample LFA or 2-PL model could be estimated, depending on whether the items are continuous or binary, with equality constraints on some or all of the item parameters. So long as there is a subset of invariant (or non-DIF) items, the mean and variance of the factor can be validly compared across studies because the factor has been placed on the same metric. Further, the factor can be treated as an outcome or predictor in other analyses with assurance that the metric of the factor is identical over studies. In some cases, such relationships can be examined directly in the model (e.g., by moving from a factor analysis to a structural equation model), whereas in other cases we may need to generate scale scores for use in a second stage analysis (e.g., when the measurement and structural models are too complex to be fit simultaneously, see e.g., Curran et al., 2008). For either purpose, the ideal is for more of the items to be invariant than not, lending greater confidence to the assertion that the latent variables are commensurately scaled across studies (Reise, Widaman & Pugh, 1993; Steenkamp & Baumgartner, 1998).

In principle, the latent factor can be put on the same scale when less than half (even just one) of the items is invariant over studies (Steenkamp & Baumgartner, 1998). In practice, however, the likelihood of correctly identifying the invariant and non-invariant items decreases as the proportion of non-invariant items increases (French & Finch, 2008; Millsap, 2005; Yoon & Millsap, 2005). If the wrong item(s) are selected as invariant, then across-study comparisons will be biased. Unfortunately, in many IDA applications, a large number of “common” items will actually be harmonized variables, and harmonization may fail to produce items that function equivalently across studies (e.g., the harmonized income items discussed previously). Hence, for IDA, we must expect to observe (and tolerate) more DIF than we would typically find if we were to administer an identical scale to distinct samples.

IDA will often pose additional challenges. Many of these challenges again arise from the use of different measurement instruments across studies. First, there will often be many non-common items unique to each instrument (and thus each study). These non-common items may be viewed as “missing” for the other studies. Fortunately, the maximum likelihood estimator permits partially missing data under the Missing at Random (MAR) assumption (Arbuckle, 1996; Wothke, 2000; see Schafer & Graham, 2002). As data for the non-common items would be missing by design (due to the use of different measurement instruments across studies), the MAR assumption would seem quite plausible. The inclusion of non-common items is thus not as problematic as might be supposed.

Second, the number of common items may in some cases be quite small. In the extreme, there might be only one common item against a field of non-common items. In this case, constraining the item parameters of the solitary common item to be equal across studies putatively puts the factor on the same scale, but there would be no opportunity to conduct DIF (invariance) tests for this (or any other) item. The advantage of multiple common items is the ability to conduct DIF tests to formally evaluate, rather than simply assume, measurement equivalence over studies.

Third, there may be no items that are common to all studies. Fortunately, this situation can be accommodated as long as the studies can be chained together in some fashion. For instance, if item sets A and B were available in Study 1, item sets B and C were available in Study 2, and item sets C and D were available in Study 3, invariance constraints on the B and C item parameters would place the factor on a common metric in all three studies. This would permit the comparison of Study 1 and Study 3, despite the fact that no common items were available for those two studies. In effect, the items administered in Study 2 enable the disjoint items sets from Studies 1 and 3 to be linked together.[3] An example of how measures can be linked across studies in this manner to facilitate IDA is described by Curran et al. (2008). They used IRT methodology to develop a commensurate measure of internalizing behavior across three longitudinal studies, one which used a subset of items from the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1981), a second which used the full CBCL in addition to the Brief Symptom Inventory (BSI; Derogatis & Spencer, 1982), and a third which used only the BSI. The second study provided information with which to equate internalizing scores across all three studies. This situation parallels the common-item design used to construct developmental scales, in which each pair of adjacent grades receives a set of common items so that test scores can be linked in a chain across all of the grades (Kolen & Brennan, 2004, p. 372-380).

Two other challenges associated with IDA are somewhat more problematic for traditional psychometric models. The first of these is that the items available for a given construct may vary in scale type (i.e., continuous, ordinal, binary, count) across or within studies. The LFA model is appropriate only for continuous items, whereas the 2-PL model is appropriate only for binary items. Similarly, most other psychometric models were explicitly developed for items that are homogenous with respect to scale type, limiting their application for IDA. Recent years, however, have seen the development of more flexible measurement models that can accommodate items of different types, and we will see that this generalized modeling framework can greatly facilitate IDA. A second challenge is that we will often anticipate DIF both across and within studies. For instance, Curran et al. (2008) examined DIF for the CBCL and BSI internalizing items over three studies, between genders, and across age groups. Given the limitations of current multi-sample modeling approaches, Curran et al. (2008) evaluated DIF across these dimensions non-simultaneously and by splitting the sample on age (10-17 years versus 18-33 years). Clearly, it would be preferable to be able to evaluate DIF across multiple variables simultaneously and without the need to bin continuous variables like age into discrete groupings. In particular, artificial categorization may greatly reduce the power to detect DIF (see MacCallum, et al. 2002, for a general discussion of problems caused by categorizing continuous variables). We focus on addressing these two additional challenges in the remainder of the paper through the development of a new, more general model, which we call Moderated Nonlinear Factor Analysis (MNLFA).

Moderated Nonlinear Factor Analysis

We shall begin by describing recently proposed Generalized Linear Factor Analysis (GLFA) models that can incorporate items of mixed scale types, including binary, ordinal, continuous and count variables (Bartholomew & Knott, 1999; Moustaki, 1996; Skrondal & Rabe-Hesketh, 2004). We will then show how the GLFA model can be extended so that observed variables like study or age can serve as moderators of the model parameters. We will continue to focus on models with a single factor to simplify the exposition, although multifactor models are also possible.

[3]If some studies share no common items with any other studies, one can conduct a “measurement study” in which all or some of the items from each of the original studies are administered to new (but similar) participants. The measurement study would then be included in the IDA to knit together the measures from the original studies.

The Generalized Linear Factor Analysis Model

Paralleling the Generalized Linear Model (GLM) for observed regressors (McCullagh & Nelder, 1989), the GLFA model may be written as

$$g_i(\mu_{ij})=\nu_i+\lambda_i\eta_j \tag{3}$$

or, equivalently, as

$$\mu_{ij}=g_i^{-1}(\nu_i+\lambda_i\eta_j) \tag{4}$$

where μ_{ij} is the expected value of item *i* for person *j*. Similar to the other models we have considered, individuals are assumed to be independent of one another, and item responses are assumed to be conditionally independent across items within individual. Like the LFA model in Equation 1, ν_i and λ_i represent the intercept and factor loading for the item and η_j is the latent factor, assumed to be normally distributed as $\eta_j \sim N(\alpha, \psi)$. Unlike Equation 1, however, the model given in Equation 3 is written in terms of the expected value, hence the absence of a residual term. Uncertainty in the actual item response is instead implicitly included in the model by specifying the (conditional) response distribution for the item, where this distribution may be chosen to be any member of the exponential family (e.g., normal, Poisson, binomial, etc). In relation to a typical GLM, $\nu_i + \lambda_i\eta_j$ takes on the role of the linear predictor, and g_i is the link function. The inverse link function is g_i^{-1}. For example, for a binary item, a natural choice for the response distribution is the Bernoulli distribution (or binomial distribution with one trial). This distribution is defined simply as $P(y_{ij} = 1|\eta_j) = \mu_{ij}$ and $P(y_{ij} = 0|\eta_j) = 1 - \mu_{ij}$. By also choosing the logit link for g_i, one obtains the usual logistic regression model, with the caveat that the single predictor is a latent factor.

For ordinal outcomes, a slight modification of Equation 3 is required because one must simultaneously model multiple category probabilities. In this case, the response distribution would be indicated as multinomial, with *C* categories. A model is then specified for $C - 1$ *cumulative* probabilities, denoted as $\mu_{cij} = P(y_{ij} \leq c)$ where $c = 1, 2,\ldots, C - 1$ (the cumulative probability for the last category must, by definition, be 1, so is not modeled). The term μ_{ij} in Equation 3 is replaced by a vector of cumulative probabilities, each of which is modeled as

$$g_i(\mu_{\mathrm{cij}})=\tau_{ci} - (\nu_i+\lambda_i\eta_j) \tag{5}$$

where τ_{ci} is a threshold parameter and the other parameters are interpreted as before. The motivation for the threshold parameter follows from thinking about the ordinal variable as a coarse representation of an underlying continuous variable. Say the ordinal item has 3 categories. When the underlying continuous variable is below τ_{1i} then the observed score is equal to 1. If it exceeds τ_{1i}, it is equal to 2 or 3. Categories 2 and 3 are, in turn, separated by τ_{2i}. The full set of thresholds and the intercept ν_i are not jointly identified. Commonly, τ_{1i} is set to zero so that ν_i (and all other thresholds) can be uniquely estimated. Alternatively, ν_i can be set to zero so that all thresholds can be estimated. In either case, choosing the logit link for g_i results in the equivalent of a cumulative logit or proportional odds model with a latent regressor.

A defining feature of the GLFA model is that Equation 3 is linear in form (as is Equation 5). Since η_j is normally distributed, ranging over the real line, this implies that $g_i(\mu_{ij})$ must also range from -∞ to ∞. But the range of expected values possible for a given item may not be infinite. The role of the link function g_i is thus to transform the (often) limited range of μ_{ij} into values that range between -∞ to ∞. Seen the other way around, the inverse link function g_i^{-1} maps the infinite range of values obtained from the linear predictor, $\nu_i + \lambda_i\eta_j$, into an interval that is appropriate for the expected value of the item. For instance, for a binary item, selecting the logit link ensures that the GLFA implies expected values (predicted probabilities) only between zero and one. Note that the link function is subscripted by *i* to indicate that this choice is item specific, as items with different response distributions will demand different link functions.

To clarify the components of the GLFA model, we shall consider how it subsumes the LFA model in Equation 1 and the 2-PL IRT model in Equation 2. Let us first assume all items are continuous. To reproduce the normal-theory LFA model, we would specify that the conditional response distribution of each item is normal, written in symbols as $y_{ij}|\eta_j \sim N(\mu_{ij}, \sigma_i^2)$. We would also choose the identity link, defined as $g_i(\mu_{ij}) = \mu_{ij}$, to relate the linear predictor to the conditional mean of each item. Equation 3 would then reduce to

$$\mu_{ij}=\nu_i+\lambda_i\eta_j \tag{6}$$

which is precisely what the LFA model in Equation 1 also implies. Thus the GLFA model in Equation 3 reduces to a conventional LFA if all items are assumed to be conditionally normal and the identity link function is selected for each item.

Now let us suppose that all items are binary. To reproduce the 2-PL model, we would chose a Bernoulli response distribution for each item, i.e., $y_{ij}|\eta_j \sim BER(\mu_{ij})$ and the logit link function, which may be defined as $g_i(\mu_{ij}) = \ln[\mu_{ij}/(1-\mu_{ij})]$. Equation 3 would then take the form

$$\ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right)=\nu_i+\lambda_i\eta_j \tag{7}$$

To see the relation of this model to Equation 2 better, we can re-express Equation 7 using the inverse link function:

$$\mu_{ij}=\frac{1}{1+\exp\left[-(\nu_i+\lambda_i\eta_j)\right]} \tag{8}$$

The terms within the exponential function in the denominator can then be rearranged to result in

$$\mu_{ij}=\frac{1}{1+\exp\left\{-\lambda_i\left[\eta_j-(-\nu_i/\lambda_i)\right]\right\}} \tag{9}$$

Comparing to Equation 2, we can see that this model is an alternative, but equivalent, parameterization of the 2-PL model where $a_i = \lambda_i$ and $b_i = -v_i/\lambda_i$ (and the notation η_j is used in place of θ_j to represent the latent factor; Takane & de Leeuw, 1987).

The real power of this generalized framework, however, is that we can select different response distributions and different link functions for different items included in the same analysis. Thus, unlike many traditional psychometric models, not all items need be of the same scale type. We can, for instance, simultaneously choose the identity link and normal distribution for a continuous item, the logit link and Bernoulli distribution for a binary item, and the log link and Poisson distribution for a count item. Ordinal items may also be included by selecting a multinomial distribution and using a logit link function for the cumulative probabilities. This generality depends on the assumption of conditional independence for the items, which permits the model to be defined by unique univariate distributions for each item, rather than a multivariate distribution for the full set of items (see Bartholomew & Knott, 1999, p. 4, and Appendix). The assumption of conditional independence is not overly restrictive, as sources of dependence can be accommodated by adding more latent factors to the model (i.e., method factors), although this can make estimation of the model more difficult.

Moderated Nonlinear Factor Analysis

The MNLFA model extends the foregoing GLFA model in several ways. First, we do not require that the response distributions for the items belong to the exponential family. One could use, for instance, a censored normal distribution for continuous subscale scores that exhibit a lower and/or upper bound. Or one might select a zero-inflated Poisson distribution for a count item which exhibits an excess of zeros. Second, although we will not explore this in depth here, we do not require a *linear* predictor – the items may be nonlinear functions of the item parameters and the latent factor (such as in a 3-PL model, see Thissen & Orlando, 2001). What is most important for the present purposes, however, is that we also allow the parameters of the model to be moderated by observed predictors, including both discrete (e.g., study) and continuous (e.g., age) variables. It is this last extension of the model, inspired in part by de Boeck & Wilson's (2004) work on “explanatory” IRT models, that offers us the opportunity to test whether our measures are commensurately scaled (i.e., invariant) when conducting IDA.

In MNLFA, the key parameters of the factor model (i.e., the indicator intercepts, factor loadings and factor mean and variance) are permitted to vary as a function of one or more exogenous moderator variables. More specifically, we express variance parameters as a log-linear function of the moderators (this prevents the predicted variance from ever being negative), and all other parameters as linear functions of the moderators (other functions could, however, be entertained). In the current context, we are particularly concerned with moderation of the item parameters, as this would indicate DIF, or a lack of measurement invariance.

Recall that when evaluating measurement invariance, either within the factor analytic or IRT modeling frameworks, it is important to first model differences in the factor mean and variance that might otherwise masquerade as DIF. Thus, the first change we make to the GLFA model described above is to allow the factor mean and variance to differ across levels of the exogenous moderators. The latent factor is thus now assumed to be distributed as $\eta_j \sim N(\alpha_j, \psi_j)$ with

$$\alpha_j = \alpha_0 + \sum_{q=1}^{Q} \alpha_q x_{qj} \tag{10}$$

$$\psi_j = \psi_0 \exp\left(\sum_{q=1}^{Q} \omega_q x_{qj}\right) \tag{11}$$

An observed moderator variable is indicated as x_q, with Q moderators in total. Although this same notation is used in each equation (and those to follow), the moderators included in the equations could potentially differ. The effect of a moderator on the factor mean is given by the term α_q, whereas the effect of a moderator on the factor variance is given by the term ω_q. The parameters α_0 and ψ_0 are the factor mean and variance of η when all moderators are equal to zero. Typically, we will fix α_0 and ψ_0 to 0 and 1, respectively, to set the scale of the latent factor and identify the model. The remaining parameters in Equations 10 and 11 may then be estimated so long as at least a subset of the items display invariant intercepts and factor loadings.

To test for non-invariant items, we also allow for moderation of the item parameters. We shall thus modify Equation 3 to be

$$g_i(\mu_{ij}) = \nu_{ij} + \lambda_{ij}\eta_j \tag{12}$$

where the additional j subscripts on ν and λ indicate that the intercepts and loadings may now differ (deterministically) across individuals as a function of the moderators. In particular, we shall express both the intercepts and loadings for the items as linear functions of the moderators as follows:

$$\nu_{ij} = \nu_{0i} + \sum_{q=1}^{Q} \nu_{qi} x_{qj} \tag{13}$$

$$\lambda_{ij} = \lambda_{0i} + \sum_{q=1}^{Q} \lambda_{qi} x_{qj} \tag{14}$$

Here, ν_{0i} and λ_{0i} are the intercept and factor loading values for item i for an individual scoring zero on all moderators. The parameters ν_{qi} and λ_{qi} indicate how the intercept and factor loading for item i change as a function of the level of the moderator x_q, holding all other moderators constant.

For ordinal items, the set of item parameters also includes thresholds (see Equation 5), which can similarly be specified as linear functions of the moderator variables:

$$\tau_{ci} = \tau_{0ci} + \sum_{q=1}^{Q} \tau_{\mathrm{cqi}} x_{qj} \tag{15}$$

Here, τ_{0ci} is the value of threshold c for item i when all moderators are zero and τ_{cqi} conveys the effect of moderator q on threshold c.

Finally, some items may be characterized by response distributions that include a variance (or scale) parameter, as is true for the normal and censored normal distributions. For such items it may also be necessary to model differences in this parameter. For instance, for an item with a conditional normal distribution, $y_{ij}|\eta_j \sim N\left(\mu_{ij}, \sigma_{ij}^2\right)$, the variance parameter may be expressed as the log-linear function

$$\sigma_{ij}^2=\sigma_{0i}^2\exp\left(\sum_{q=1}^{Q}\delta_{qi}x_{qj}\right) \tag{16}$$

The parameter σ_{0i}^2 indicates the unique variance of the item when all moderators are zero; exp (δ_{qi}) is the factor by which the unique variance increases for each unit change in x_q, holding all other moderators constant.

Given the complexity of these equations, it is helpful to consider how they relate to the more traditional multi-sample LFA and 2-PL models. Suppose that there is a single binary moderator, x_1, indicating the group (e.g., study) membership of each individual. The reference group is coded $x_1 = 0$ and the focal group is coded $x_1 = 1$. Additionally, the latent factor is scaled by setting $\alpha_0 = 0$ and $\psi_0 = 1$. In this situation, Equations 10 and 11 reduce to

$$\alpha_j=\alpha_1 x_{1j} \tag{17}$$

$$\psi_j=\exp\left(\omega_1 x_{1j}\right) \tag{18}$$

For the reference group (i.e., when $x_1 = 0$), the factor mean and variance are then simply 0 and 1 (as exp (0) = 1). For the focal group (i.e., when $x_1 = 1$), the factor mean is α_1 and the factor variance is exp(ω_1). For instance, if $\omega_1 = .7$, then the factor variance in the focal group is exp (.7) = 2, or twice as large as the factor variance in the reference group. Note that this parameterization is directly equivalent to the parameterization recommended by Yoon & Millsap (2007) and Embretson & Reise (2000, p. 252-262) for fitting multi-sample linear factor analysis and 2-PL IRT models, respectively, in which the factor mean and variance are set to 0 and 1 for one group, but freely estimated in the other.

Similarly, when the item parameters are moderated by a single binary predictor Equations 13 and 14 reduce to

$$\nu_{ij}=\nu_{0i}+\nu_{1i}x_{1j} \tag{19}$$

$$\lambda_{ij}=\lambda_{0i}+\lambda_{1i}x_{1j} \tag{20}$$

Here, ν_{0i} and λ_{0i} are the intercept and factor loading for item *i* within the reference group. For the focal group, the intercept and factor loading are $\nu_{0i} + \nu_{1i}$ and $\lambda_{0i} + \lambda_{1i}$, respectively. Thus, λ_{1i} and λ_{1i} represent *differences* in the intercepts and factor loadings across groups (i.e., DIF). In multi-sample factor analysis or IRT, it is more conventional to directly estimate the intercepts

and factor loadings for the two groups, rather than their differences, but the two approaches yield equivalent results.

Finally, if all items are assumed to have conditionally normal response distributions, as in the linear factor analysis, then we would also need to account for differences in the unique variances of the items across groups via Equation 16. With the binary moderator, Equation 16 simplifies to

$$\sigma_{ij}^2=\sigma_{0i}^2\exp\left(\delta_{1i}x_{1j}\right) \tag{21}$$

where σ_{0i}^2 is the conditional variance of the item for the reference group, and $\sigma_{0i}^2\exp(\delta_{1I})$ is the conditional variance of the item for the focal group. Thus, $\exp(\delta_{1i})$ is the ratio of the conditional variances for the two groups, where a value of exp(0) = 1 would indicate equal variances. Again, this parameterization differs from the LFA model, in which the conditional variances for the two groups are usually estimated directly, rather than their ratio, but the results are equivalent.

Thus, Equations 10, 11, 13, 14, and 16 can be specified to fully replicate a multi-sample LFA or 2-PL model if the only moderator is a grouping variable. More generally, however, these same equations can accommodate multiple moderators which may be categorical and/or continuous. We must recognize, however, that there will often be practical limitations to the amount of moderation that can be assessed given the extent of information available in the data. The evaluation of DIF should therefore be motivated by specific theoretical and/or methodological concerns. For instance, in our empirical demonstration of IDA, we will simultaneously evaluate DIF as a function of study and chronological age (measured continuously). DIF across studies is expected due to concerns with harmonized items, whereas DIF by age is anticipated due to changes in alcohol use norms over the long period of development under consideration.

Example: Measuring Alcohol Involvement

As children enter adolescence and then progress to adulthood, rapid and important changes take place in their involvement with alcohol. Here, we investigate these changes using data from two independent longitudinal studies from age 10 to 22. We begin by briefly describing each study sample. We next discuss the items available for measuring alcohol involvement within each study and then proceed through the analysis. We describe both a general model fitting strategy and the generation of scale scores for later analyses.

Samples

For each study, a primary objective was to contrast the development of children of alcoholics (COAs) to matched control participants (children of non-alcoholics). The first study is the Adolescent/Adult Family Development Project (AFDP; Chassin, Rogosch & Barrera, 1991), which includes 246 COAs and 208 controls. AFDP includes data from three annual interviews conducted when the target child was an adolescent (10-16 years of age at Wave 1, 1988) and a young-adult follow-up interview conducted approximately 5 years after Wave 3 (age 17-23). The second study is the Alcohol and Health Behavior Project (AHBP; Sher, Walitzer, Wood & Brent, 1991). In this study, 489 college freshmen (250 COAs and 237 controls) completed four annual assessments beginning in 1987. The age span represented in the AFDP study permits us to examine how alcohol involvement changes over early adolescence through emerging adulthood. Supplementing this data with the AHBP data allows us to examine consistency in the pattern of change identified in early adulthood and augments our ability to

detect effects in this important developmental period. Table 1 reports the number of participants observed at each age in each study.

Since the MNLFA model requires conditional independence of the response vectors across participants, we randomly selected a single time point of data per participant (despite the availability of repeated measures) to create a "calibration sample" for establishing the measurement properties of the alcohol involvement factor. In fitting models to the calibration sample, a key goal is to obtain valid item parameter estimates that can be used to generate scale scores for the full set of observations. This strategy was also used by Curran et al. (2008) to ensure independence across participants when estimating 2-PL models to obtain measures of internalizing behavior for IDA. In total, our calibration sample included observations from 454 and 484 participants from AFDP and AHBP, respectively, as shown in Table 1. After fitting the MNLFA models to the calibration sample, scoring for the alcohol involvement factor was carried out for the entire set of 3602 observations (1728 from AFDP and 1874 from AHBP) to permit future analyses (e.g., growth curve modeling of alcohol involvement) to include all repeated measures (see, e.g., Curran et al., 2008).

Indicators of Alcohol Involvement

We consider five indicators of alcohol involvement, listed in Table 2 in decreasing order of severity. *Disorder* is an indicator of the presence or absence of alcohol abuse or dependence, as determined by the Diagnostic Interview Schedule (DIS-III-R) in the AFDP sample or the DIS-III in the AHBP sample. *Consequences* is a count of the number of both consequences of alcohol use and symptoms of alcohol dependence experienced over the past year by the participant (e.g., got in trouble at school or work, got into a physical fight, destroyed property, etc), with a maximum number of nine. Response scales differed within and between studies for specific consequences and symptoms, so these were collapsed into binary values and summed to produce the overall count. *Heavy Use* is an ordinal item indicating how often participants consumed 5 or more alcoholic drinks within the past year. In AFDP, the original items were scored from 0 to 7, whereas in AHBP, they were scored from 0 to 9, with somewhat different category labels in each study. Due to sparseness in the endorsement of some response options, and to make the heavy drinking items comparable across studies, the responses were harmonized into three categories, 0 = not in the past year, 1 = less than once a month, and 2 = once a month or more. The *Use Frequency* indicator was based on similar items, except with reference to any use of alcohol. The original items were again harmonized into three categories, 0 = not in the past year, 1 = less than once a month, and 2 = once a month or more. Finally, *Expectancies* is a measure of positive alcohol expectancies and was computed as the mean of four items scored 0 to 4 in each of the two studies (e.g., drinking alcohol helps me when I feel tense or nervous, helps me celebrate social occasions, etc.). The wording of the items and anchor labels differed slightly between studies.

The indicators of alcohol involvement culled from the two studies thus differ in their response scales: one is dichotomous, one is a count, two are ordinal, and another is a bounded continuous variable. This necessitates use of a model which will allow for alternative response distributions and link functions for the indicators. Additionally, although we have attempted to make the items comparable in a face valid way through harmonization, it is unclear whether the items will in fact function equivalently across studies. Alcohol use norms also change with age, potentially altering the significance of certain item responses (e.g., frequency of use, given increasing acceptability and availability of alcohol in young adulthood relative to early adolescence). For these reasons, we need to allow the parameters of the model to be moderated by study and age, necessitating the use of an MNLFA model. No other model can address all of these issues simultaneously; only the MNLFA enables the IDA to move forward.

Measuring and Modeling Alcohol Involvement

Our analysis of alcohol involvement from age 10 to 22 involved four steps. First, we conducted a nonlinear factor analysis that pooled across the two studies (assuming all parameters to be equal over groups, i.e., without moderation). Second, we expanded this model by permitting the factor mean and variance to vary both as a function of (continuous) age and (discrete) study. Third, we evaluated possible DIF over study and age by simultaneously testing whether these two variables moderated the intercepts or factor loadings of the indicators. Fourth, we generated alcohol involvement scores across all timepoints based on the estimates obtained from the calibration sample. The results from each step are described below in turn. All models were fit using the NLMIXED procedure in SAS 9.1 using adaptive quadrature with fifteen quadrature points (see Appendix for details on estimation and scoring). Documentation on how to arrange the data, fit the models, and graph the results is provided at <address to be determined>. Although we do not provide the details here, we also conducted a small-scale simulation study based on our sample characteristics to confirm that NLMIXED would provide unbiased estimates for our MNLFA mods when using 15 quadrature points.

Nonlinear Factor Analysis of Alcohol Involvement—We posited a one-factor model for the five indicators, shown in path diagram form in Figure 2. Within the diagram, a single-headed arrow indicates that the variable pointed to is regressed on the variable from whence the arrow originated, whereas a double headed arrow indicates a variance or residual variance parameter. For the current set of indicators, only Expectancies has a response distribution with a variance parameter, hence it is the only indicator shown with a double-headed arrow. Means and intercepts are implicit but not depicted. The latent variable is labeled as η and is interpreted to represent alcohol involvement. To identify the model, the mean and variance of η were set to zero and one, respectively. All item intercepts and loadings could then be estimated. Table 2 summarizes the response distributions and link functions selected for each item.

The estimates obtained from this initial model are presented in Table 3. Of primary interest are the factor loadings, all of which were positive and statistically significant. This implies that as a participant's level of η goes up, they have increasingly positive alcohol expectances, use alcohol more frequently, engage in heavy drinking more often, report more adverse consequences of alcohol use, and have a higher probability of meeting criteria for abuse or dependence. This pattern of results is consistent with the interpretation of η as an 'alcohol involvement' factor. Relative comparisons of the item parameters are somewhat uninformative, however, given the different scale types and link functions of the items. A better understanding of how the indicators relate to the factor can be obtained by graphing the relationships, similar to the ICC or traceline curves conventionally used in IRT analyses. For the binary and ordinal items, Figure 3 plots the probabilities of the item responses as a function of η. For the count and continuous items, the expected value of the indicator is plotted against the value of η. These plots were constructed by selecting values of η between -2.5 and 2.5 (in standard deviation units) and using these in combination with the estimated item parameters in the inverse link function to obtain the corresponding predicted values for μ for each item (see Equation 4).

Examining the plot for Disorder, we can see that the probability of meeting criteria for abuse or dependence is very low for values of η below the mean. The probability of disorder increases rapidly as η exceeds the mean, reaching .56 when η is one standard deviation above the mean, and about .92 when η is two standard deviations above the mean. Similarly, for Consequences, we can see the expected number of consequences or symptoms is roughly zero for values of η below the mean. One consequence is predicted for individuals who are one standard deviation above the mean on η, and approximately four consequences are predicted for individuals who are two standard deviations above the mean. Turning to our two ordinal items, the three lines

in these panels trace out the probability that the indicator will be scored 0, 1, and 2, respectively, as a function of η. For Heavy Use, we can see that individuals below the mean on η are most likely to report no heavy alcohol use in the past year, individuals just above the mean on η are most likely to report heavy use less than once a month, and individuals with high levels of η are most likely to report heavy use more than once a month. For Use Frequency, the curves are similar, but shifted to the left. This shift simply indicates that participants will engage in occasional or frequent alcohol use at lower levels of η than they will engage in occasional or frequent *heavy* alcohol use. Finally, in the lower left of Figure 3 we see the relation between η and positive alcohol expectancies. The plot is slightly nonlinear given the lower censoring at zero. For uncensored observations, the average Expectancy score increases by about .5 units (on a 0 to 4 scale) for each standard deviation increase in η.

Age and Study Differences in Alcohol Involvement—The analysis presented above incorporated no information about study differences or age differences in alcohol involvement. To gain a better sense of what these effects might be, we conducted a preliminary graphical analysis. We first computed the observed means for each indicator at each year for the two studies. These means were then transformed by the same link functions listed in Table 2 for the indicators and plotted as a function of Age and Study in Figure 4.[4] The purpose of applying the transformations was to put the graphs into the scale of the linear predictor for each item (i.e., Equation 3). Figure 4 provides us with two important pieces of information. First, it shows a consistent developmental pattern in which rapid increases are observed until about age 17 or 18, after which little further change is observed. The uniformity of this trend is consistent with our assumption that all of these variables are reflections of a single underlying factor, supporting the validity of the model. Second, consistent differences across age or between studies are indicative of factor mean shifts, whereas minor inconsistencies suggest possible age or study DIF. Here, study differences are less pronounced and less consistent than developmental differences. In particular, AHBP participants report more consequences of alcohol use and higher frequencies of alcohol use and heavy use, but have less positive alcohol expectancies. The two studies are similar on disorder rates. The inconsistency of the study differences suggests that the indicators may not be functioning equivalently across the two studies, a finding that is not suprising given the use of harmonized indicators.

To model overall age and study differences in alcohol involvement, we fit the MNLFA model depicted in Figure 5. The moderation of the factor mean and variance by Age and Study is indicated by arrows to the factor and factor variance, respectively. Based on the plots in Figure 4, we elected to model age-related changes in mean levels of alcohol involvement with a piecewise linear model (Raudenbush & Bryk, 2002, pp. 178-179). This entailed constructing a new age variable, labeled “Age17” which was scored as Age in Years −17 for participants 17 and under, and simply scored zero for participants older than age 17. Study is coded 0 for AFDP and 1 for AHBP. The model for the factor mean was then specified as

$$\alpha_j = 0 + \alpha_1 \mathrm{Age17}_j + \alpha_2 \mathrm{Study}_j$$

Note that the intercept of this equation, α_0, was set to zero to identify the factor mean, implying that the factor mean is zero from age 17 onwards in the AFDP study. We did not include a second slope to allow for mean changes after age 17, as is sometimes done in piecewise linear models, because the plots in Figure 4 suggested relative stability from 17 to 22 years of age.

[4]Note that at some ages the transformations were undefined (e.g., a log of zero), in which case no points are plotted in Figure 4 for those particular ages. Additionally, age 17 was not plotted for the AHBP study, given that there were only two participants of that age in the calibration sample.

Similarly, the relative uniformity of the developmental trends observed in Figure 4 suggested no need for an interaction between age and study.

The model for the factor variance was specified via the log-linear equation

$$\psi_j = 1 \exp(\omega_1 \text{Age17}_j + \omega_2 \text{Study}_j)$$

The parameter ω_1 allows for continuous changes in the variance of the latent factor from age 10 to age 17, with stability assumed thereafter, and the parameter ω_2 allows for study differences in the variance of the factor. The model was identified by setting the ψ_0 coefficient to 1, implying that the variance of the factor in the AFDP study is 1 from age 17 to age 22.

Overall, the inclusion of these four effects in the model resulted in a large improvement in model fit, as indicated by the likelihood ratio test (χ^2(df=4) = 4066.10, $p < .0001$). There was a significant age effect on the factor mean ($\hat{\alpha_1} = .43$, SE=.05, $p < .0001$), but the age effect on the factor variance was not significant ($\hat{\omega_1} = -.03$, SE = .07, $p = .66$). We also detected significant study differences in the factor mean ($\hat{\alpha_2} = .48$, SE = .09, $p < .0001$) and variance ($\hat{\omega_2} = -.47$, SE = .17, $p < .01$). As seen in Figure 6, the study differences indicated that the overall level of alcohol involvement is higher in the AHBP sample than the AFDP sample, whereas the variance is somewhat smaller. This pattern of results may be due to the fact that the AHBP sample is college-based, whereas AFDP is community based. To better determine the validity of these mean and variance differences across age and study, we next evaluated the assumption that the indicators are functioning identically for all participants.

Evaluating Measurement Invariance—We employed a stepwise procedure to evaluate DIF, similar to the strategies described earlier for LFA and 2-PL models. Both Age and Study DIF were considered, with Age DIF limited to the span between 10 and 17 years of age.[5] We first examined the improvement in the log-likelihood that was obtained by allowing for Age and Study DIF in a single indicator, considering each indicator in turn. If allowing for DIF resulted in a significant improvement in model fit for more than one indicator then we determined the indicator for which DIF most improved the model fit. Allowing for DIF in this indicator, we then examined the additional improvement in model fit that would accrue by allowing for DIF with a second indicator, considering each remaining indicator in turn. This process was repeated until no significant improvement in model fit was obtained by allowing for DIF in any additional indicator. The total possible number of likelihood ratio tests involved in evaluating DIF in this way for five indicators is 14, so a Bonferroni-corrected alpha level of .004 was used to select the indicators with significant DIF (alternatively, we could have used the Benjamini-Hochberg procedure to control the false discovery rate; Thissen, Steinberg & Kuang, 2002;Thissen, Steinberg & Wainer, 1993).

Allowing for Age and Study DIF in the intercept and loading of the Use Frequency indicator resulted in the largest improvement in model fit (χ^2 (4) = 129, p < .0001). Retaining the DIF parameters for Use Frequency, we then tested each remaining indicator for DIF. The largest improvement in model fit resulted from including Age and Study effects on the intercept, loading and residual variance of the Expectancies indicator (χ^2 (6) = 102, p < .0001). Also allowing for DIF in the Disorder indicator further improved the fit of the model (χ^2(4) = 32.1, p < .0001). The model with DIF for Use Frequency, Expectancies, and Disorder could not be

[5]This choice reflected our belief that DIF would be most likely to occur during periods of rapid developmental change. Such mean changes are not, however, necessary to observe DIF. As an alternative, we could have modeled DIF by specifying the intercepts and loadings as linear or quadratic functions of continuous age. Additionally, given the current specification (using the Age17 variable and partialling study DIF) and the age ranges within the two studies, the age DIF obtained here pertains only to AFDP.

further improved by allowing for DIF in either Heavy Use ($\chi^2(4) = 5.0$, p = .29) or Consequences ($\chi^2(4) = 8.1$, p = .09). The intercepts and loadings of these two indicators are thus considered to be invariant to Age and Study. Heavy Use and Consequences thereby form the most stable components of the alcohol involvement latent factor. For the other three indicators, we removed DIF parameters that were estimated as non-significant to arrive at a final model. Removing these parameters did not result in a significant decrement in model fit ($\chi^2(7) = 9.0$, p = .25). The final, reduced model is shown in Figure 7. Moderation of the indicator intercept is shown as a direct arrow from the moderator to the indicator, whereas moderation of the factor loading shown as an arrow pointing to the arrow originating from η.

For Use Frequency, all four DIF parameter estimates were significantly different from zero. Both the intercept and factor loading for this indicator increased from age 10 to age 17. The meaning of these effects can be seen visually in Figure 8. Each panel of Figure 8 displays the predicted response probabilities (tracelines) for the Use Frequency categories for a particular age and study. Comparing across the first three panels, we can see that, holding constant η, older individuals are more likely to endorse the 1 or 2 categories of this indicator. This reflects the fact that, for young adolescents (e.g., age 10), drinking at all (even less than once a month) is indicative of a relatively high level of alcohol involvement. For older adolescents and young adults, engaging in occasional or even frequent alcohol use is more normative and does not necessarily indicate a high level of overall alcohol involvement. In addition to these age effects, there were also significant positive effects of study on the intercept and loading of the Use Frequency indicator. These effects are seen in the comparison of the bottom two panels of Figure 8. As can be seen, in AHBP, the tracelines are shifted to the left, indicating that the 1 and 2 category responses are observed at lower levels of η in this sample than AFDP. One possible explanation for why alcohol use is more common at lower levels of η in the AHBP sample is that it is a college-based sample (whereas AFDP is not) for which higher levels of alcohol use may be more normative.

For Expectancies, there was a significant study effect on the intercept, such that positive alcohol expectancies were a half-point lower in the AHBP sample than the AFDP sample (on a scale of 0 to 4) at any given level of η. This effect is consistent with our earlier observation that, in Figure 4, only Expectancies showed a lower mean level in the AHBP sample than the AFDP sample (all other indicators have comparable or higher mean levels in AHBP). The intercept difference for Expectancies absorbs this discrepancy. Additionally, the residual variance was negatively related to age. From age 10 to age 17, the standard deviation of the residuals decreased from 1.8 to .8. One interpretation of this effect is that Expectancies become a more reliable indicator of alcohol involvement as adolescents increase in age and enter adulthood.

Finally, there was a significant positive effect of age on the intercept for Disorder. As shown in Figure 9, the probability of meeting criteria for abuse or dependence at a given level of η is lower for younger adolescents than older adolescents or young adults. One possible explanation for this finding is that diagnostic interviews based on DSM criteria for adults are insensitive to alcohol disorder in adolescents (Martin et. al, 1995, 1996).

Allowing for DIF for these three items resulted in a stronger effect of Age on the factor mean ($\hat{\alpha_1}$ = .50, SE=.06, $p < .0001$) and, more so, the factor variance ($\hat{\omega_1}$ = −.12, SE = .06, $p = .053$). The factor mean difference across studies diminished ($\hat{\alpha_2}$ = .30, SE = .10, $p < .01$), whereas the factor variance difference increased ($\hat{\omega_2}$ = −.57, SE = .19; $p < .01$). These new estimates account for the fact that some indicators are non-equivalent (show DIF) across Age and Study.

Scoring—We next computed MAP scores for the alcohol involvement factor. Such scores are particularly useful when subsequent models cannot realistically be estimated simultaneously with the measurement model of the MNLFA, such as growth curve modeling

(e.g., Hussong et al. 2007, 2008). Of course, in using scores one must take care to evaluate their quality. To this end, we plotted the standard errors of the alcohol involvement MAPs for the calibration sample in Figure 10. Notice that the standard errors are much higher for low scores. This pattern reflects the fact that the indicators included in our model do not provide much information on low levels of alcohol involvement. Returning briefly to Figure 3, we can see that most of the "action" in the item plots takes place around or above the mean of η, with comparatively little information to discriminate between individuals much below the mean. The sole exception is the plot for the Expectancies indicator, for which the expected value steadily increases as a function of alcohol involvement. Indeed, it is precisely for this reason that we included the Expectancies indicator in the model – to better differentiate between individuals with relatively low levels of alcohol involvement, who had perhaps not yet begun drinking alcohol. Nevertheless, we can see that the majority of the information provided by the indicators remains at the high end of the scale, as evidenced by the lower standard errors at and above about -1 in Figure 10.[6]

For comparison, we also computed MAPs from our preceding two models, a nonlinear factor analysis without age or study effects, and a nonlinear factor analysis where only the factor mean and variance were conditional on age and study (but no DIF). Our primary purpose was to determine which model extensions had a meaningful impact on the scores. As can be seen in Figure 11, the inclusion of age and study effects on the factor mean and variance resulted in noticeably different MAPs than those produced from a nonlinear factor analysis model without age or study effects, especially for individuals with low levels of alcohol involement. Tuning the scores to the specific age and study of the participant thus appears to have been quite important. In contrast, the additional inclusion of age and study DIF had much less influence on the scores. The fact that DIF was relatively inconsequential for scoring in the present analysis provides some additional confidence that we are in fact measuring the same construct in each study at each age. In other applications, the impact of DIF on the scores may be more marked.

Using the parameter estimates obtained from the full model (including DIF), we next generated MAPs for all of the repeated measures (not just those included in the calibration sample). Figure 12 shows boxplots of the MAP scores obtained at each age in each study, with white boxes indicating AFDP and gray boxes indicating AHBP. Superimposed on the boxplots are the means (bold lines) ± two standard deviations (light lines) implied by the final MNLFA fit to the calibration sample (akin to Figure 6, but now incorporating DIF). As can be seen, the boxplots follow the mean trends quite closely. As implied by the model, there is an overall linear increase in alcohol involvement from age 10 to 17, after which the mean level of alcohol involvement is stable through age 22. Further, a slightly higher level of involvement is observed in the strictly college-based AHBP sample than the community-based AFDP sample.

In contrast to the mean trends, however, the dispersion of the MAP estimates disagrees with the model-implied pattern. The boxplots indicate *increasing* dispersion with age, whereas the model implies *decreasing* dispersion with age. The reason for this disagreement relates to the point made above, that there is less information for estimating low alcohol involvement scores. As noted previously, when items provide less information, as occurs here at lower levels of alcohol involvement, MAPs are shrunken more towards the marginal means (i.e., the MAPs are pulled towards the bold lines in Figure 12). Since alcohol involvement increases with age, there is less shrinkage at older ages (where scores are relatively high) than younger ages (where

[6]Note that the scale of η is not the same in Figures 3 and 10. Figure 3 was generated from a model in which η was scaled as a standard normal, with mean 0 and standard deviation 1. The scores shown in Figure 10 were generated from a model in which η was scaled to be standard normal for 17 year olds in the AFDP study. Nevertheless, making relative comparisons between the two figures can be helpful for seeing why the standard errors are higher in Figure 10 for lower estimates.

scores are relatively low). When MAPs are plotted across age, as in Figure 12, there is then the appearance of increasing variance when, in fact, individual differences in alcohol involvement are narrowing. The skew seen in the MAP distributions observed at the younger ages is likely of the same origin. In general, one must be mindful that factor score estimates may not always have the same properties as the true scores (Grice, 2001), potentially distorting the results of subsequent analyses, such as growth models. To mitigate the problem, one can try to identify additional indicator variables to improve score estimation and fill in gaps in information (i.e., provide greater information about individual differences at the low end of the scale), however we do not pursue this further here.

Conclusions, Limitation, and Directions for Future Research

A fundamental challenge of IDA is that we require commensurate measures, but independently conducted studies often measure constructs using different items, inventories, or even modes of assessment. Harmonization provides an ad hoc and face valid way to score items similarly, but this alone does not ensure that the item scores have the same meaning with respect to the underlying construct. Traditional and novel psychometric models offer an opportunity to formally test this assumption, and to relax it for a subset of items as necessary. Procedures for testing whether items function similarly across samples are well-developed for both the linear factor analysis model and other commonly used IRT models. Where these procedures may prove overly restrictive, however, is in their requirement that each item or indicator of the latent factor be of the same scale type (e.g., all continuous or all binary) and their inability to evaluate measurement equivalence across both categorical and continuous variables. The proposed MNLFA model overcomes both of these limitations. Our application of this model involved a combination of binary, ordinal, count, and bounded-continuous items, and simultaneously allowed for DIF due to Study and Age. It is difficult to imagine another strategy that could have been used to combine these indicators into a single, commensurately scaled summary measure. Additionally, it is worth noting that, although our example included only common items, items unique to each study can also be included in an MNLFA and this may result in different numbers of items for different studies.

One assumption that remains restrictive, however, in both the MNLFA as well as more traditional psychometric models, is that the sample units are independent of one another. It was this assumption that led us to sample one record per person from the AFDP and AHBP studies to create our analysis sample, despite the availability of repeated measures for each participant. Clearly, the use of all available observations would be preferable, as this would improve estimation. It would also allow the MNLFA to be expanded from a measurement model to a model of greater substantive interest. For instance, we could model individual trajectories of change in alcohol involvement directly within the MNLFA rather than through subsequent analyses of MAP scores (similar to a second-order growth model; Sayer & Cumsille, 2001, Johnson & Raudenbush, 2006). Unfortunately, evaluating the likelihood function of the model via deterministic methods of numerical integration like quadrature becomes increasingly intractable as the number of random effects needed to account for dependence due to clustering increases. The development and application of improved stochastic methods of estimation may help to overcome this limitation in the future and should be the focus of continued methodological research (Cai, 2008; Fox, 2003; Fox & Glas, 2001).

More broadly, the problem of measurement can also be minimized at the design stage. Investigators embarking on new studies should take care to include measures used in similar studies, either in whole or in part, to facilitate possible future IDA. No harmonization will be necessary for the identical items, likely resulting in fewer items displaying DIF. With a higher ratio of non-DIF to DIF items, we can have greater confidence that the measurement is truly

commensurate over studies, strengthening the IDA. In general, a coordinated effort to adopt standard measures across studies would greatly benefit IDA.

In sum, the problem of measurement, although fundamental for IDA, can be solved through the careful application of psychometric models. Further, the continued development and refinement of these models will offer new opportunities for IDA in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the National Science Foundation (award SES-0716555, D. Bauer), and the National Institute on Drug Abuse (award R01 DA15398, A. Hussong). The studies that supplied data for our integrative data analysis example were supported by the National Institutes on Alcohol Abuse and Alcoholism (awards R01 AA16213, L. Chassin and R01 AA013987, K. Sher) and Drug Abuse (award R01 DA05227, L. Chassin). We are grateful to the Statistical and Applied Mathematical Sciences Institute (SAMSI), Research Triangle Park, NC, for sponsoring a program on Latent Variable Models in the Social Sciences in 2004-2005, during which time the initial development of the moderated nonlinear factor analysis model presented here began. We also thank Li Cai for helpful discussions and assistance with the empirical analyses, and Sonya Sterba and Nisha Gottfredson for their research on factor score estimation in nonlinear factor analysis models.

Appendix: Estimation and Scoring for the MNLFA Model

The full MNLFA model exceeds the current estimation capabilities of most latnet variable modeling software. At present, the more restrictive GLFA model can be fit via maximum likelihood estimation with several software packages, including GLLAMM (Rabe-Hesketh, Pickles & Skrondal, 2001) and Mplus (Muthén & Muthén, 2007). A large number of MNLFA models can also be fit by these programs if the desired response distributions are included among the software options and the linear form of Equation 12 is adopted. Moderation of the factor mean and item intercepts can be accomplished within these programs by regressing the factor and items, respectively, on the exogenous variables (see Muthén, 1989). In addition, factor loading differences can potentially be modeled in Mplus by including interaction terms between the moderators and the latent factor in the model. Allowing for moderation of variance parameters and thresholds by exogenous variables is, however, somewhat more challenging within current software. To our knowledge, the multi-sample framework is the only available option within these programs which allows for the moderation of *all* types of parameters, and this approach requires a single categorical moderator variable to define the samples. Thus the full MNLFA model cannot be fit using currently available latent variable modeling software.

Fortunately, as we now consider in detail, estimation of the full MNLFA can be carried out in SAS by capitalizing on the fundamental similarity between nonlinear mixed effects models and nonlinear factor analysis. Let $\boldsymbol{\gamma}$ be a vector of all model parameters to be estimated. $\boldsymbol{\gamma}$ is composed of two subvectors: $\boldsymbol{\pi}$, parameters governing the distribution of the latent factor(s) (i.e., parameters in Equations 10 and 11), and $\boldsymbol{\varpi}$, parameters governing the conditional item distributions (i.e., parameters in Equations 13-16). Further, let $\mathbf{y}_j$ be a vector of item responses and $\mathbf{x}_j$ be a vector of values for the exogenous moderator variables for sample unit *j*. The marginal likelihood for the model may be written as

$$L(\gamma)=\prod_{j=1}^{N}\int\varphi(\eta_j|\mathrm{x}_j,\pi)\prod_{i=1}^{P}f_i(y_{ij}|\eta_j,\mathrm{x}_j,\varpi)d\eta_j \tag{A1}$$

where $\varphi(\boldsymbol{\eta}_j|\mathbf{x}_j, \boldsymbol{\varpi})$ is the (normal) density function for the latent factor(s) and $f_i(y_{ij}\backslash\boldsymbol{\eta}_j; \mathbf{x}_j, \boldsymbol{\varphi})$ is the (conditional) univariate response distribution for item *i*. Note that the first product operator in Equation A1 follows from the assumption that the item response vectors are independent across sample units after conditioning on $\mathbf{x}_j$. Similarly, the second product operator in Equation A1 follows from the assumption that, within sample unit, the item responses are independent after conditioning on $\boldsymbol{\eta}_j$ and $\mathbf{x}_j$. This second assumption permits the specification of distinct (conditional) univariate density functions for each item, allowing for items with different scale types (e.g., Bernoulli for binary, Poisson for count, etc).

The likelihood function in Equation A1 is of the same form as that optimized by the SAS NLMIXED procedure (SAS Institute, 2008, p. 4375-4376). The NLMIXED procedure was originally developed for fitting nonlinear mixed effects models. Such models are designed for clustered data structures in which multiple observations are taken within each Independent Sampling Unit (ISU), e.g., multiple individuals within each of many groups, or repeated measures on each of many individuals. Naturally, observations taken on the same ISUs are expected to be more similar to one another than observations taken on different ISUs. This dependence of observations due to unobserved heterogeneity across ISUs is accounted for by including random (ISU-specific) effects in the model. For instance, a random intercept term accounts for general (unexplained) differences in the level of the dependent variable across ISUs. Customarily, the random effects are assumed to be normally distributed, and the observations within clusters are assumed to be independent, conditioning on the random effects.

Parallels between mixed effects models (both linear and nonlinear) and latent factor/trait models have been noted by many authors (see Bauer, 2003; de Boeck & Wilson, 2004; Curran, 2003; Mehta & Neale, 2005; Rijmen et al., 2003; Skrondal & Rabe-Hesketh, 2005). For a typical latent variable model, item responses may be construed as “clustered” within participant. The dependence of the item responses within participants is assumed to arise through individual differences on one or more latent factors, which may be viewed as the random effects of the model. Accounting for the latent factors, the item responses are assumed to be conditionally independent, just as observations are assumed to be independent within clusters after conditioning on the random effects. Hence, regarding *j* as ISU, item responses $i = 1, 2, \ldots, P$ within *j* as clustered observations within ISUs, and latent factors as random effects, Equation A1 equally describes the marginal likelihood for a nonlinear mixed effects model. Given the correspondence between the two models, the NLMIXED procedure can be used to fit the MNLFA model, minimizing $-\log L(\boldsymbol{\gamma})$ to arrive at the parameter estimates.

To compute the likelihood, the NLMIXED procedure implements numerical methods to approximate the integral in Equation A1, including fixed-point Gauss-Hermite quadrature and adaptive Gaussian quadrature. These approaches involve approximating the integral as a weighted sum over pre-defined locations on the distribution of $\boldsymbol{\eta}$. The quality of this approximation depends on the number of locations (i.e., quadrature points) and their placement with respect to the distribution of $\boldsymbol{\eta}$. Non-adaptive quadrature selects the locations for all units with respect to the marginal (“prior”) distribution of $\boldsymbol{\eta}$, and the locations are fixed across successive iterations. If the locations are not well selected initially (which may occur as a consequence of supplying poor start values to the algorithm) then fixed-point quadrature can yield inferior estimates (see Skrondal & Rabe-Hesketh, pp. 167-169; Molenberghs & Verbeke, 2004). In contrast, adaptive quadrature selects the locations at each iteration with respect to the posterior distribution of $\boldsymbol{\eta}$) given the data and current estimates of $\boldsymbol{\gamma}$. More specifically, the points are centered around the posterior mode. Additional details on the NLMIXED procedure, including quadrature and optimization algorithms, may be obtained from SAS Institute (2008, Chapter 61).

To compute the posterior mode, either for estimation or scoring purposes, NLMIXED makes use of the following version of Bayes' Theorem:

$$f(\eta_j|\mathrm{y}_j, \mathrm{x}_j, \widehat{\gamma}) \propto \varphi(\eta_j|\mathrm{x}_j, \widehat{\pi}) \prod_{i=1}^{P} f_i(y_{ij}|\eta_j, \mathrm{x}_j, \widehat{\varpi}) \tag{22}$$

where $f(\boldsymbol{\eta}_j|\boldsymbol{y}_j; \mathbf{x}_j; \boldsymbol{\gamma}\hat{})$ is the posterior distribution of $\boldsymbol{\eta}_j$ given the vector of parameter estimates $\boldsymbol{\gamma}\hat{}$ and the observed data for unit *j*. The posterior is proportional to the product of the normal "prior" for the factor and joint likelihood of the item responses. Thus, minimizing the negative log of this product returns the mode of the posterior distribution for unit *j* (SAS Institute, 2008, p. 4376). This value is referred to in the IRT literature as the Maximum a Posteriori (MAP) estimate of $\boldsymbol{\eta}_j$ (Bock & Aitken, 1981). With adaptive quadrature, this value is updated across iterations for the selection of the quadrature point locations. Regardless of the method of quadrature, however, MAPs can be computed for all units based on the final parameter estimates to serve as scale scores for the factors (i.e., factor score estimates).

One subtle difference between Equation A1 and the typical form of the marginal likelihood for a nonlinear mixed effects model bears special mention, namely the subscripting of f_i, the density function for the item. As described previously, this subscripting is necessary to permit the density function to differ across binary, ordinal, continuous and count items. More typically, nonlinear mixed effects models are fit to multiple observations on the same dependent variable within ISUs, and a common distribution is assumed for all observations. For instance, observations might be repeated measures on a binary variable, in which case a Bernoulli distribution could be assumed at all occasions. Fortunately, the different item distributions of the MNLFA model can be accommodated by the NLMIXED procedure through the use of programming statements to toggle between user-defined log-likelihoods for the different items. Practically, one includes IF-THEN statements in the program to indicate IF the response is to Item 1, THEN the log-likelihood is computed from the user-defined function log(f_1), IF the response is to Item 2, THEN the log-likelihood is computed from the user-defined function log (f_2), etc. Example code demonstrating the use of these features of the NLMIXED procedure for fitting and scoring the Alcohol Involvement models presented in this paper is available at <address to be determined>.

References

Achenbach T, Edelbrock C. The classification of child psychopathology: A review and analysis of empirical efforts. Psychological Bulletin 1981;85:1275–1301. [PubMed: 366649]

Angoff, WH. Perspectives on differential item functioning methodology. In: Holland, PW.; Wainer, H., editors. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993. p. 3-23.

Arbuckle, JL. Full information estimation in the presence of incomplete data. In: Marcoulides, GA.; Schumacker, RE., editors. Advanced structural equation modeling: Issues and techniques. Hillsdale, NJ: Erlbaum; 1996. p. 243-277.

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

Bauer DJ. Estimating multilevel linear models as structural equation models. Journal of Educational and Behavioral Statistics 2003;28:135–167.

Bock, RD. Different DIFs: Comment on the papers read by Neil Dorans and David Thissen. In: Holland, PW.; Wainer, H., editors. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993. p. 115-122.

Bock RD, Aitken M. Marginal maximum likelihood estimation of item parameters. an application of an EM algorithm. Psychometrika 1981;46:443–459.

de Boeck, P.; Wilson, M. Explanatory item response models: A generalized linear and nonlinear modeling approach. New York: Springer-Verlag; 2004.

Byrne BM, Shavelson RJ, Muthen B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin 1989;105:456–466.

Cai, L. Unpublished doctoral dissertation. Department of Psychology, University of North Carolina at Chapel Hill; 2008. A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement nodel.

Chassin L, Rogosch F, Barrera M. Substance use and symptomatology among adolescent children of alcoholics. Journal of Abnormal Psychology 1991;100:449–463. [PubMed: 1757658]

Cheung GW, Rensvold RB. Testing factorial invariance across groups: a reconceptualization and proposed new method. Journal of Management 1998;25:1–27.

Curran PJ. Have multilevel models been structural equation models all along? Multivariate Behavioral Research 2003;38:529–569.

Curran PJ, Hussong AM, Cai L, Huang W, Chassin L, Sher KJ, Zucker RA. Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. Developmental Psychology 2008;44:365–380. [PubMed: 18331129]

Derogatis, LR.; Spencer, P. Brief symptom inventory (BSI). Baltimore: Clinical Psychometric Research; 1982.

D'Orazio, M.; Di Zio, M.; Scanu, M. Statistical matching: theory and practice. Hoboken, NJ: Wiley; 2006.

Fox JP. Stochastic EM for estimating the parameters of a multilevel IRT model. British Journal of Mathematical and Statistical Psychology 2003;56:65–81. [PubMed: 12803822]

Fox JP, Glas CAW. Bayesian estimation of a multilevel I RT model using Gibbs sampling. Psychometrika 2001;66:269–286.

French BF, Finch WH. Multigroup confirmatory factor analysis: locating the invariant referent sets. Structural Equation Modeling 2008;15:96–113.

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

Grice JW. Computing and evaluating factor scores. Psychological Methods 2001;6:430–450. [PubMed: 11778682]

Holland, PW. A framework and history for score linking. In: Dorans, NJ.; Pommerich, M.; Holland, PW., editors. Linking and Aligning Scores and Scales. New York, NY: Springer; 2007. p. 5-30.

Holland, PW.; Dorans, NJ. Linking and equating. In: Brennan, RL., editor. Educational Measurement. Vol. Fourth. Westport, CT: American Council on Education and Praeger Publishers; 2006. p. 187-220.

Hussong AM, Flora DB, Curran PJ, Chassin LA, Zucker RA. Defining risk heterogeneity for internalizing symptoms among children of alcoholic parents: A prospective cross-study analysis. Development and Psychopathology 2008;20:165–193. [PubMed: 18211733]

Hussong AM, Wirth RJ, Edwards MC, Curran PJ, Chassin LA, Zucker RA. Externalizing symptoms among children of alcoholic parents: Entry points for an antisocial pathway to alcoholism. Journal of Abnormal Psychology 2007;116:529–542. [PubMed: 17696709]

Johnson, C.; Raudenbush, SW. A repeated measures, multilevel Rasch model with application to self-reported criminal behavior. In: Bergeman, CS.; Boker, SM., editors. The Notre Dame Series on Quantitative Methodology, Volume 1: Methodological Issues in Aging Research. Mahwah, NJ: Lawrence Erlbaum Associates; 2006. p. 131-164.

Jöreskog KG. Simultaneous factor analysis in several populations. Psychometrika 1971;36:409–426.

Kamata A, Bauer DJ. A note on the relationship between factor analytic and item response theory models. Structural Equation Modeling: A Multidisciplinary Journal 2008;15:136–153.

Kolen, MJ.; Brennan, RL. Test equating, scaling, and linking: Methods and practices. Vol. 2nd. New York: Springer; 2004.

Lazarsfeld, PF.; Henry, NW. Latent Structure Analysis. Boston: Houghton Mifflin Co.; 1968.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychological Methods 2002;7:19–40. [PubMed: 11928888]

Martin CS, Kaczynski NA, Maisto SA, Bukstein OM, Moss HB. Patterns of DSM-IV alcohol abuse and depedence in adolescent drinkers. Journal of Studies on Alcohol 1995;56:672–680. [PubMed: 8558899]

Martin C, Chung T, Kirisci L, Lagenbucher J. Item response theory analysis of diagnostic criteria for alcohol and cannibus use disorders in adolescents: implications for DSM-V. Journal of Abnormal Psychology 2006;115:807–814. [PubMed: 17100538]

McCullagh, P.; Nelder, JA. Generalized linear models. Vol. 2nd. Boca Raton, FL: Chapman & Hall/CRC Press; 1989.

Mehta PD, Neale MC. People are variables too: multilevel structural equations modeling. Psychological Methods 2005;10:259–284. [PubMed: 16221028]

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika 1993;58:525–543.

Millsap, RE. Four unresolved problems in studies of factorial invariance. In: Maydeu-Olivares, A.; McArdle, JJ., editors. Contemporary Psychometrics. Mahwah, NJ: Lawrence Erlbaum Associates; 2005. p. 153-171.

Millsap, RE.; Meredith, W. Factorial invariance: historical perspectives and new problems. In: Cudeck, R.; MacCallum, RC., editors. Factor Analysis at 100: Historical Developments and Future Directions. Mahwah, NJ: Lawrence Erlbaum Associates; 2007.

Millsap RE, Yun-Tein J. Assessing factorial invariance in ordered-categorical measures. Multivariate Behavioral Research 2004;39:479–515.

Mohlenberghs, G.; Verbeke, G. An introduction to (generalized) (non)linear mixed models. In: de Boeck, P.; Wilson, M., editors. Explanatory item response models: A generalized linear and nonlinear modeling approach. New York: Springer-Verlag; 2004. p. 111-153.

Moustaki I. A latent trait and a latent class model for mixed observed variables. British Journal of Mathematical and Statistical Psychology 1996;49:313–334.

Muthén BO. Latent variable modeling in heterogeneous populations. Psychometrika 1989;54:557–585.

Muthén, LK.; Muthén, BO. Mplus user's guide. Vol. 5th. Los Angeles, CA: Muthén & Muthén; 2007.

Rabe-Hesketh S, Pickles A, Skrondal A. GLLAMM: A generalclass of multilevel models and a Stata program. Multilevel Modelling Newsletter 2001;13:17–23.

Raudenbush, SW.; Bryk, AS. Hierarchical Linear Models: Applications and Data Analysis Methods. Vol. Second. Thousand Oaks: Sage Publications; 2002.

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin 1993;114:552–566. [PubMed: 8272470]

Rijmen F, Tuerlinckx F, de Boeck P, Kuppens P. A nonlinear mixed model framework for item response theory. Psychological Methods 2003;8:185–205. [PubMed: 12924814]

Rivers DC, Meade AW, Fuller WL. Examining question and context effects in organization survey data using item response theory. Organizational Research Methods. in press

SAS Insitute. SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute; 2008.

Sayer, AG.; Cumsille, PE. Second-order latent growth models. In: Collins, LM.; Sayer, AG., editors. New Methods for the Analysis of Change. Washington, DC: American Psychological Association; 2001. p. 179-200.

Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods 2002;7:147–177. [PubMed: 12090408]

Sher KJ, Walitzer KS, Wood PK, Brent EE. Characteristics of children of alcoholics: Putative risk factors, substance use and abuse, and psychopathology. Journal of Abnormal Psychology 1991;100:427–448. [PubMed: 1757657]

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

Sörbom D. A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology 1974;27:229–239.

Steenkamp JBEM, Baumgartner H. Assessing measurement invariance in cross-national consumer research. Journal of Consumer Research 1998;25:78–90.

Steinberg, L. When response scale labels influence the processes of self-report. Meeting of the Society for Multivariate Experimental Psychology; Chapel Hill, NC. 2007 Oct.

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

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

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

Thissen, D.; Steinberg, L.; Wainer, H. Detection of differential item functioning using the parameters of item response models. In: Holland, PW.; Wainer, H., editors. Difierential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993. p. 67-113.

Widaman, KF.; Reise, SP. Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In: Bryant, KJ.; Windle, M.; West, SG., editors. The science of prevention: Methodological advances from alcohol and substance abuse research. Washington, DC: American Psychological Association; 1997. p. 281-324.

Wirth RJ, Edwards MC. Item factor analysis: Current approaches and future directions. Psychological Methods 2007;12:58–79. [PubMed: 17402812]

Wothke, W. Longitudinal and multigroup modeling with missing data. In: Little, TD.; Schnabel, KU.; Baumert, L., editors. Modeling longitudinal and multilevel data. Mahwah, NJ: Lawrence Erlbaum Associates; 2000. p. 219-240.

Yoon M, Millsap RE. Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. Structural Equation Modeling 2007;14:435–463.

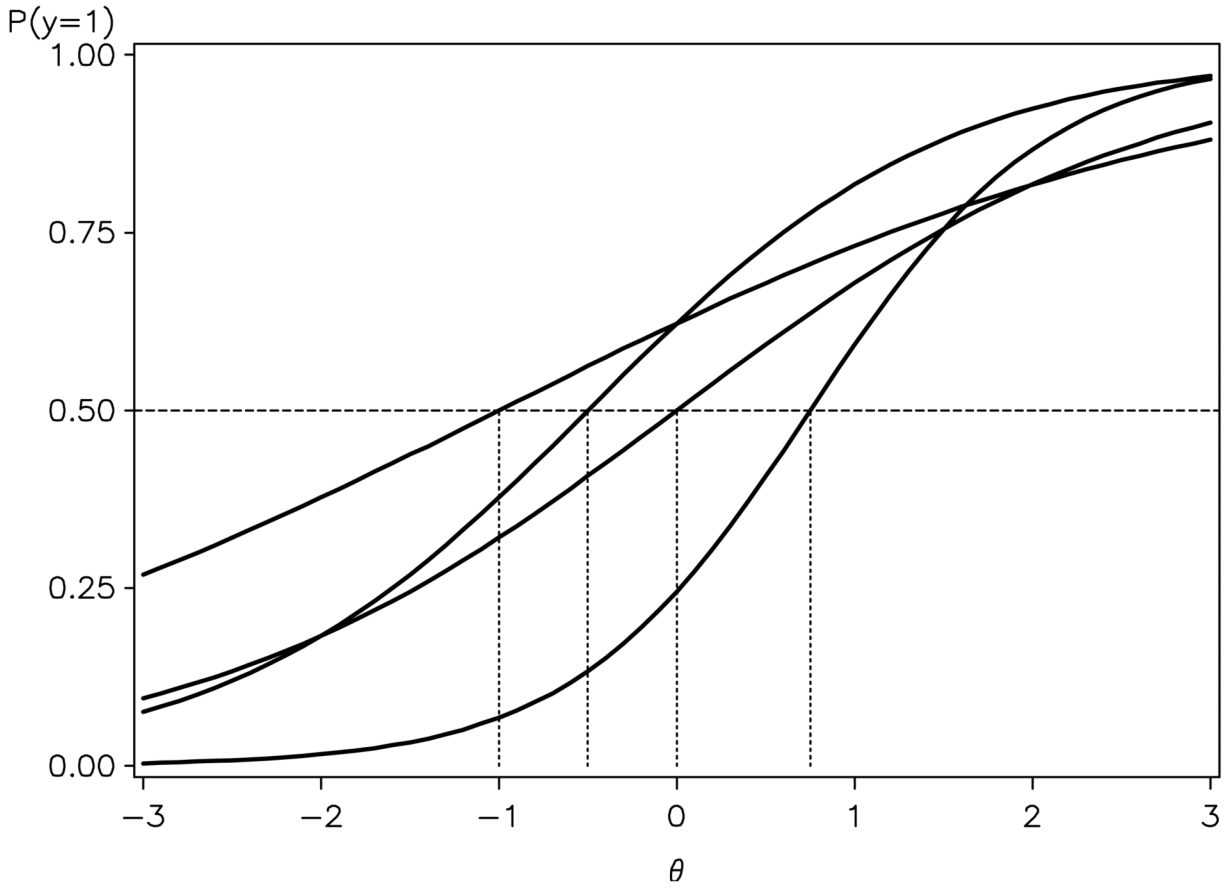

Figure 1.
Tracelines of four items in a 2-PL. Vertical droplines indicate item difficulties culties with values, from left to right, of -1, -.5, 0, and .75, with corresponding discriminations of .5, 1, . 75, and 1.5.

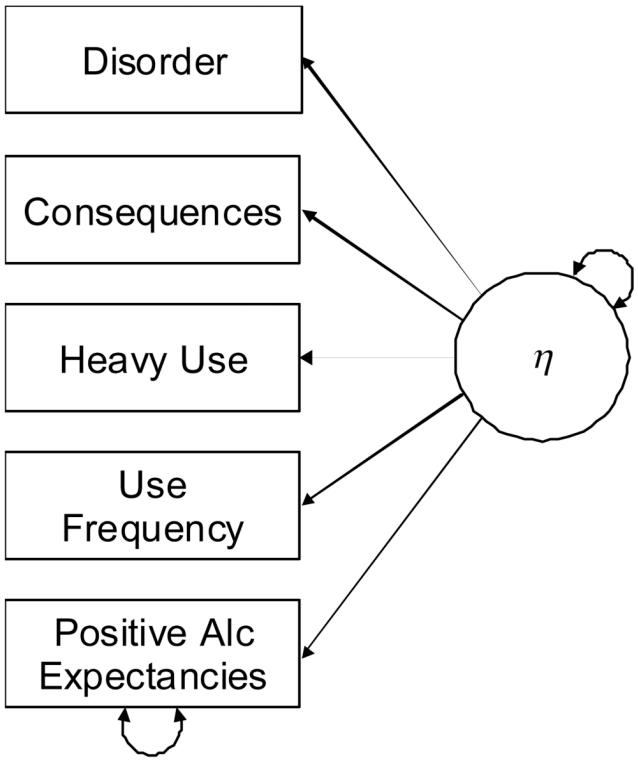

Figure 2.
Path diagram of nonlinear factor analysis model where the factor η measures alcohol involvement.

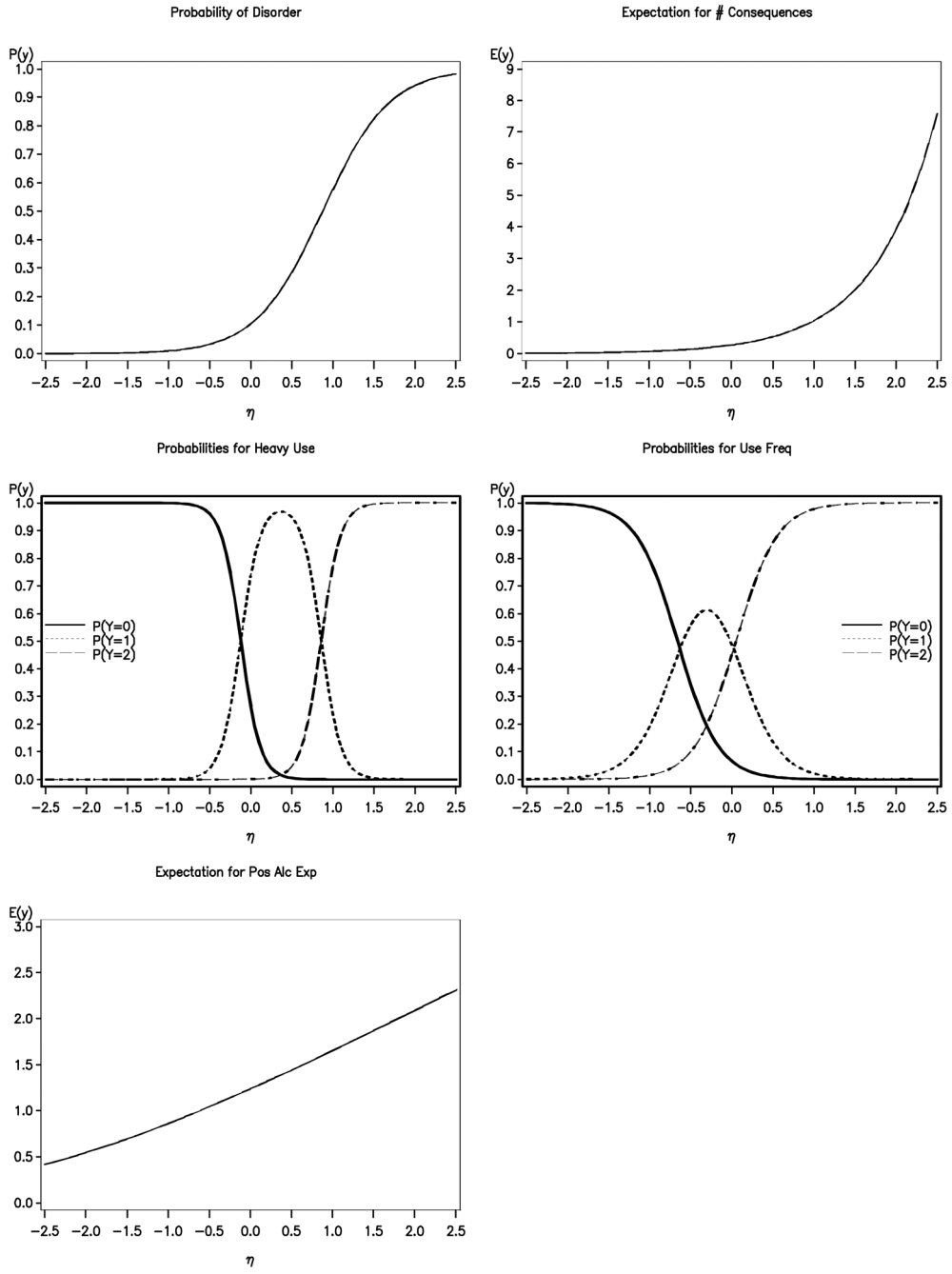

Figure 3.
Relationships between each measured variable ***y*** and the latent alcohol involvement factor η.

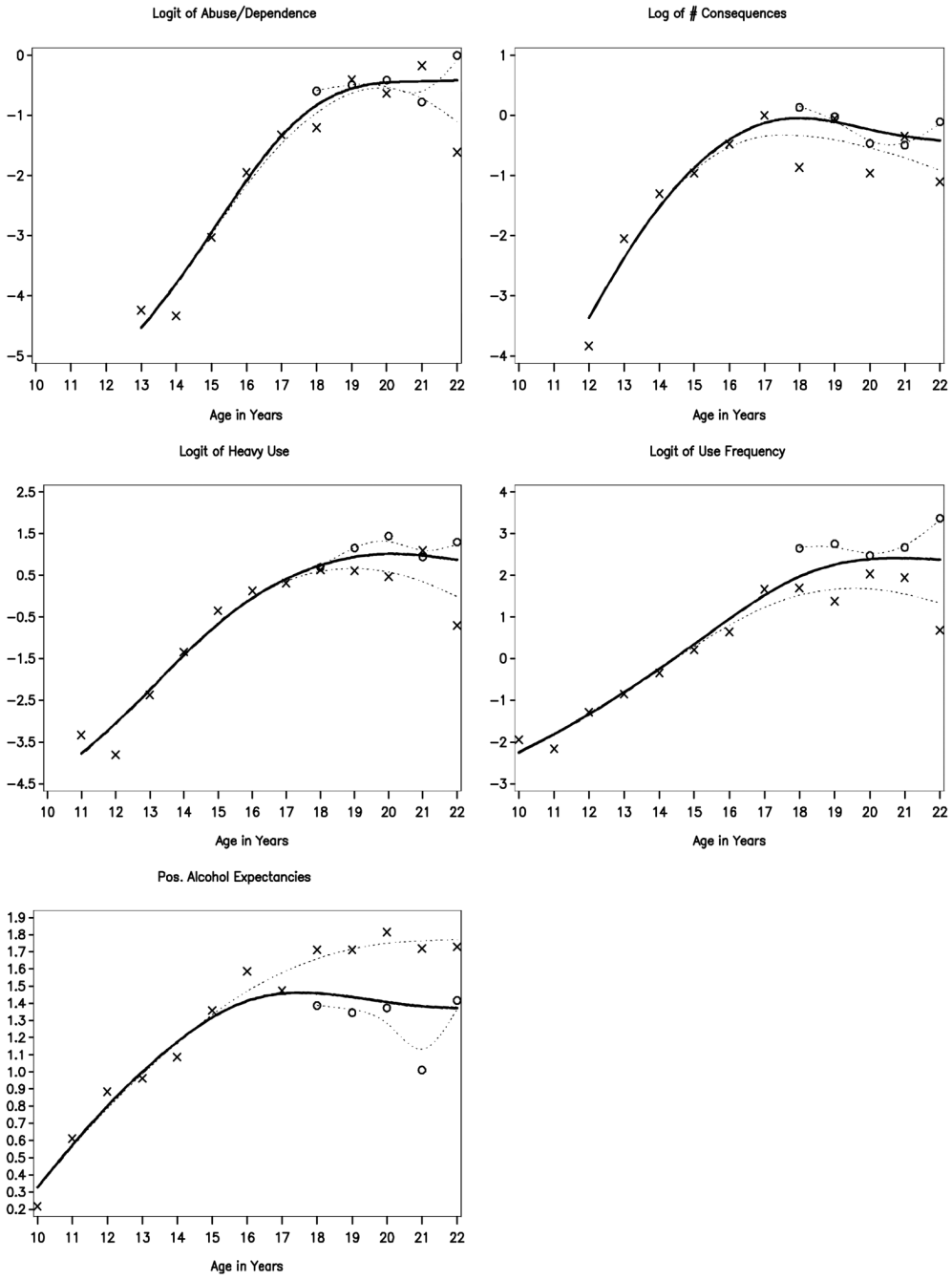

Figure 4.
Observed proportions/means of the indicator variables, submitted to link function transformation, plotted as a function of Age and Study (× = AFDP; ○ = AHBP); plots for Use Frequency and Heavy Use are based on proportions scoring 1 or 2 versus 0.

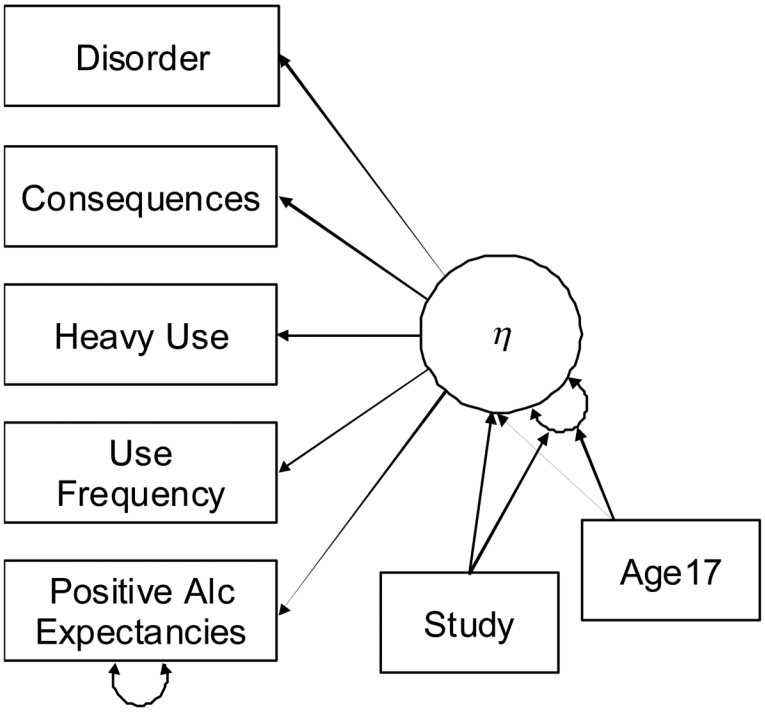

Figure 5.
Path diagram for moderated nonlinear factor analysis model with age and study effects on the mean and variance of the Alcohol Involvement factor, η.

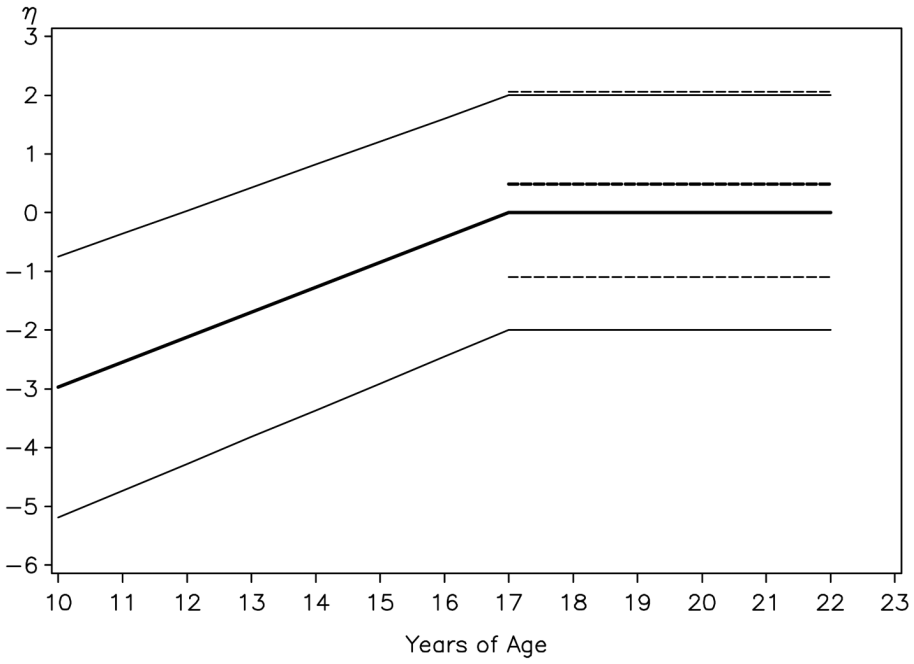

Figure 6.
Estimated Age and Study differences in Alcohol Involvement: Bold lines indicate means, light lines indicate ± 2 standard deviations; solid lines indicate AFDP and dashed lines indicate AHBP.

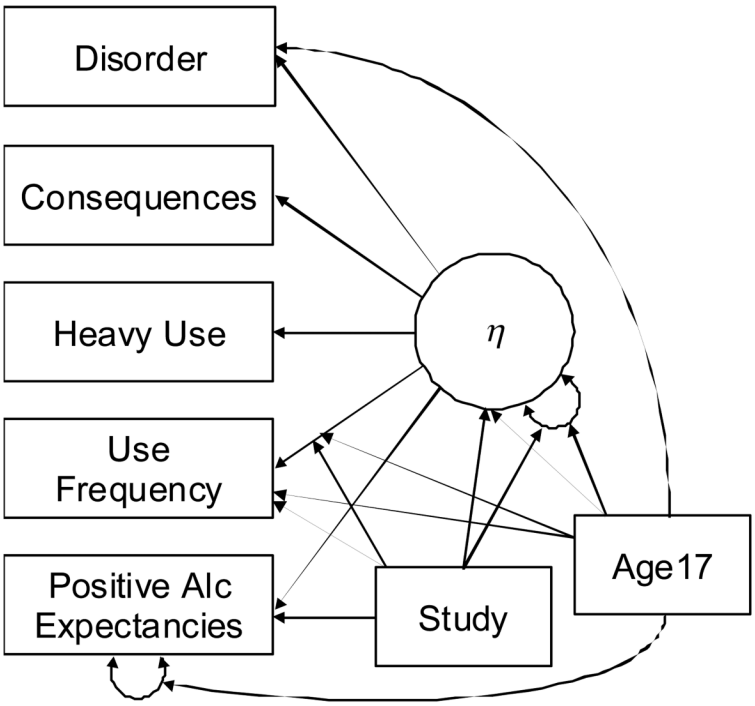

Figure 7.
Path diagram for moderated nonlinear factor analysis model with Age and Study effects on the factor mean and variance and on item parameters for indicators displaying DIF.

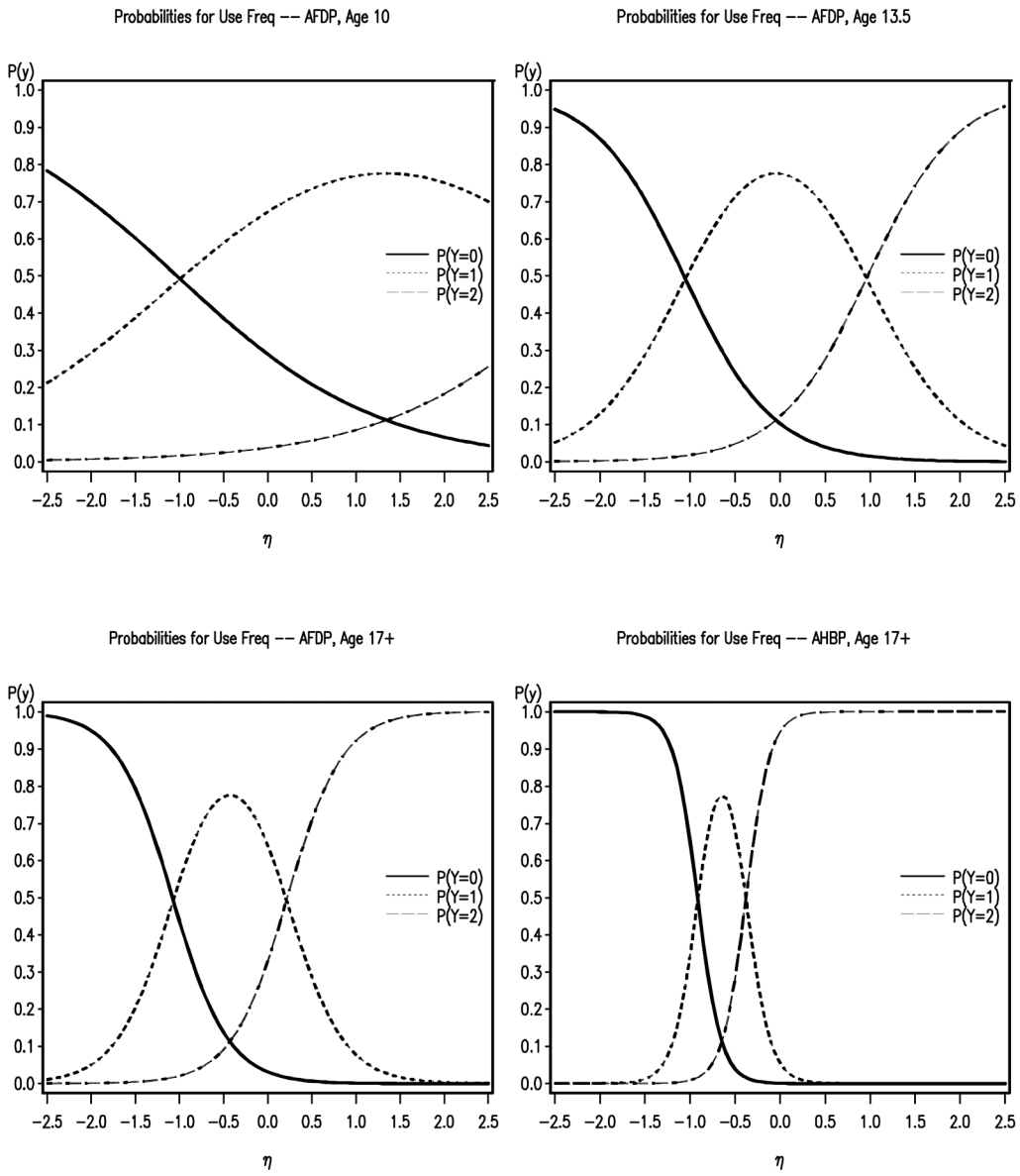

Figure 8.
Tracelines for the Frequency of Use indicator demonstrating DIF as a function of Age and Study.

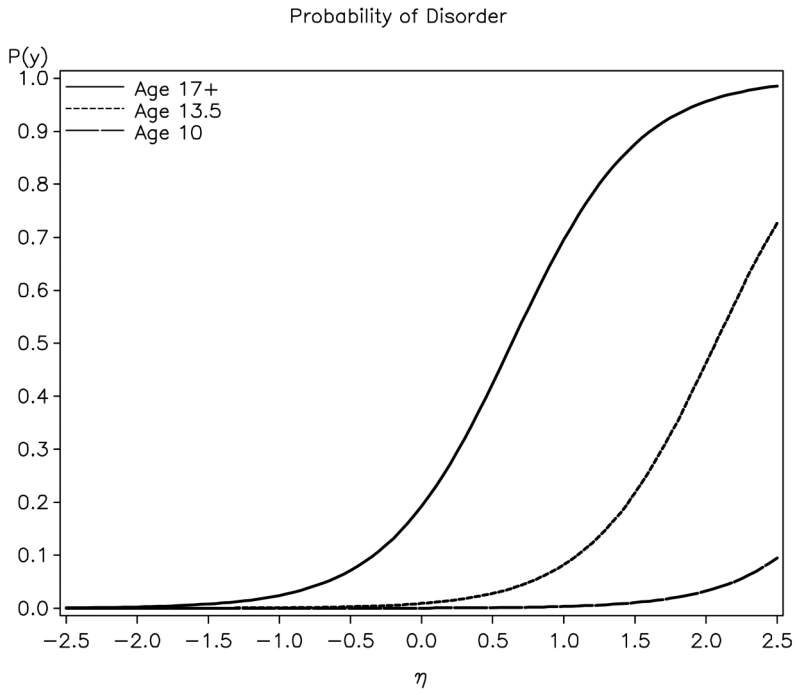

Figure 9.
Tracelines for the Disorder indicator showing DIF by Age.

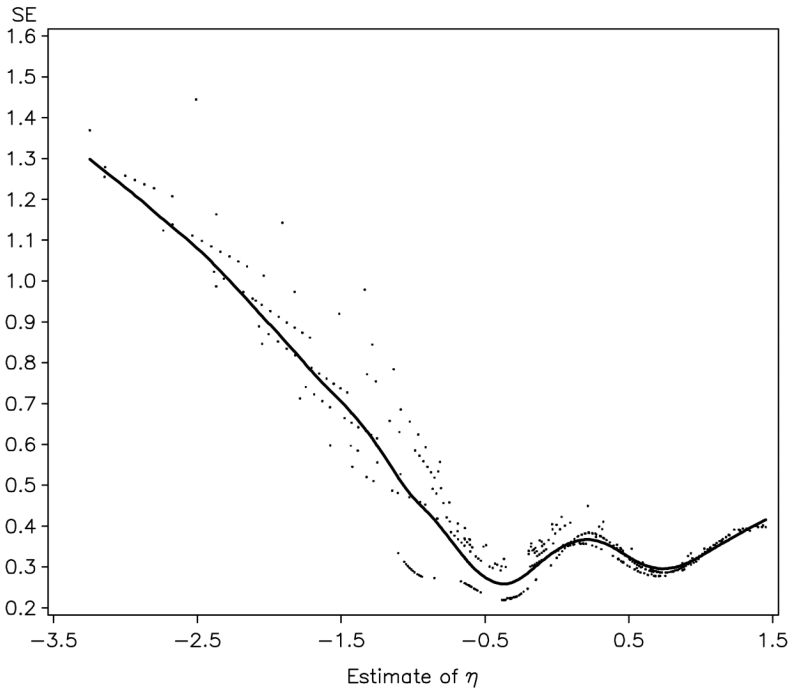

Figure 10.
Standard errors for MAP factor score estimates obtained from the MNLFA model with DIF, as a function of the estimated factor score value.

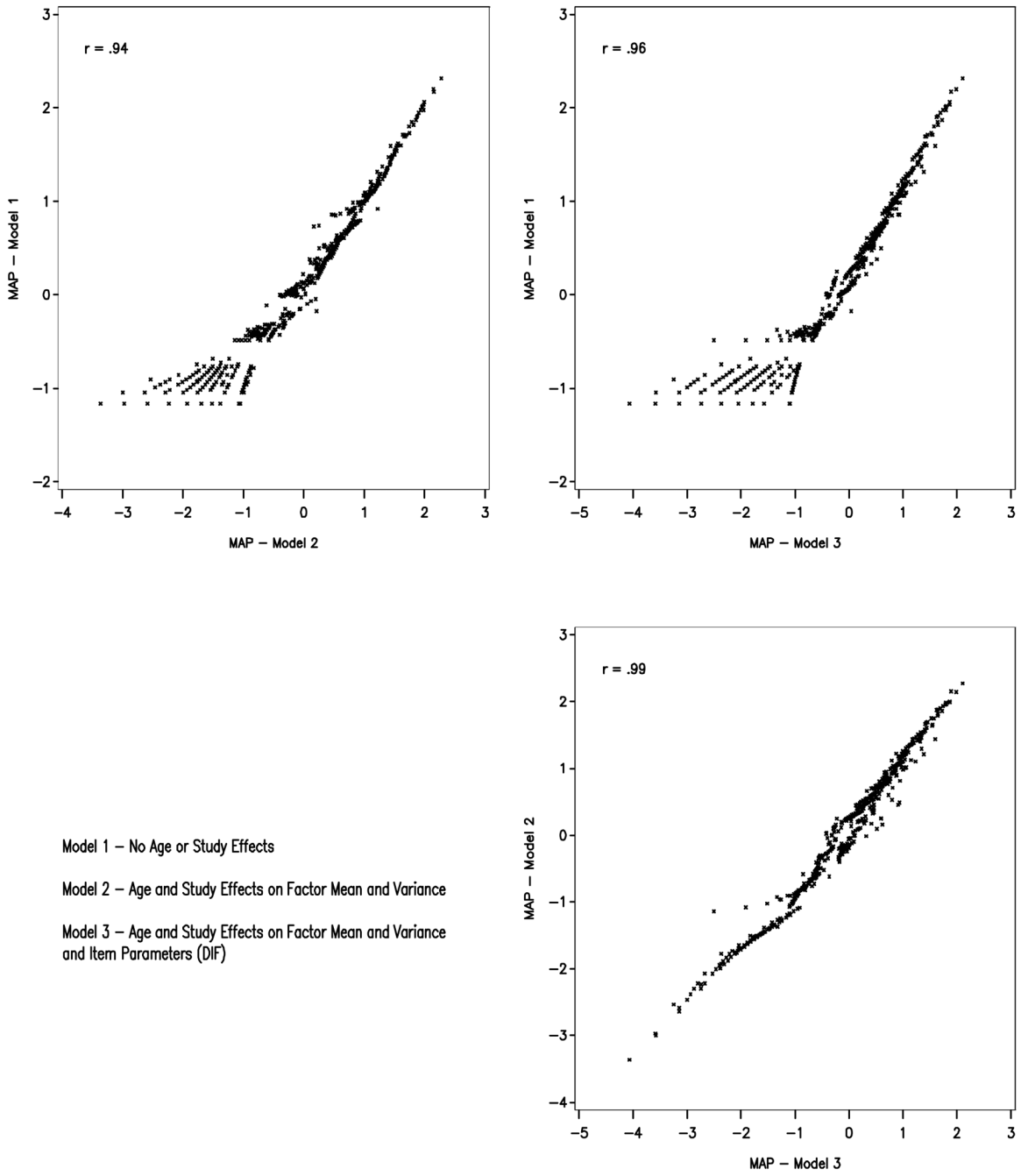

Figure 11.
MAP estimates obtained from models with and without age and study effects on the factor mean and variance and item parameters (DIF).

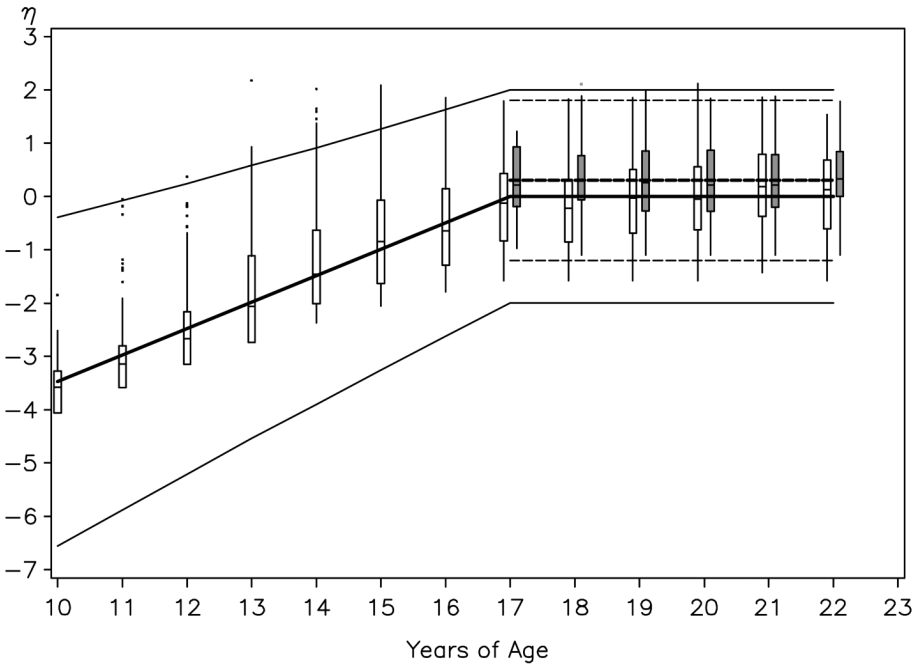

Figure 12.
Boxplots of MAP factor scores obtained at each age for the two studies, shown in relation to the model implied means (bold lines) and ± two standard deviations (light lines): AFDP indicated with white boxes and solid lines, AHBP indicated with gray boxes and dashed lines.

Table 1

Frequency of Observations in Total Sample and Calibration Sample by Age and Study Origin

	Total Sample													
	Age in Years													
Study	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**Total**
AFDP	32	106	190	264	289	244	148	57	79	80	101	96	42	1728
AHBP								8	404	472	438	434	118	1874
Total	32	106	190	264	289	244	148	65	483	552	539	530	160	3602
	Calibration Sample													
	Age in Years													
Study	10	11	12	13	14	15	16	17	18	19	20	21	22	Total
AFDP	8	29	46	70	77	65	32	19	26	20	26	24	12	454
AHBP								2	107	119	118	108	30	484
Total	8	29	46	70	77	65	32	21	133	139	144	132	42	938

Table 2

Alcohol Involvement Indicators, in Putative Order of Severity and Specifications for Moderated Nonlinear Factor Analysis (MNLFA).

			MNLFA Specifications	
Item	**Description**	**Scoring**	**Conditional Distribution**	**Link Function**
Disorder	Meets criteria for alcohol abuse or dependence	0 = absence of disorder; 1 = presence of disorder	Bernoulli	Logit
Consequences	Number of consequences or symptoms of alcohol use experienced in the past year	Count from 0-9	Poisson	Log
Heavy Use	Frequency of heavy drinking (5+ drinks at a sitting, or drunkenness) in the past year	0 = not in past year; 1 = less than once per month; 2 = once per month or more	Multinomial	Cumulative Logit
Use Frequency	Frequency of drinking in the past year (alcoholic beverage of any kind)	0 = not in past year; 1 = less than once per month; 2 = once per month or more	Multinomial	Cumulative Logit
Expectancies	Positive expectations regarding the effects of alcohol	An average of four five-point items, ranging from 0 to 4. Higher numbers indicate more positive expectations.	Censored Normal (censored from below at 0)	Linear (identity) where uncensored

Table 3

Estimates (standard errors) from non-linear factor analysis model

Item	Intercept	Loading	Thresholds 1	Thresholds 2	Residual Variance
Disorder	-2.15 (.19)	2.46 (.24)			
Consequences	-1.29 (.09)	1.33 (.08)			
Heavy Use	1.02 (.47)	8.46 (2.76)0		8.29 (2.57)	
Use Frequency	2.65 (.27)	3.99 (.38)0		2.86 (.26)	
Positive Expectancies	1.19 (.04)	0.45 (.04)			0.93 (.05)

Note. All estimates significantly different from zero at $p < .05$.