Item Response Theory

Li Cai, Kilchan Choi, Mark Hansen,
and Lauren Harrell

CRESST, University of California, Los Angeles, California 90095-1521;
email: lcai@ucla.edu

Annu. Rev. Stat. Appl. 2016. 3:297–321

First published online as a Review in Advance on March 2, 2016

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-041715-033702

Keywords

discrete multivariate analysis, item factor analysis, multilevel IRT, multidimensional IRT, diagnostic classification models, model fit testing

Abstract

This review introduces classical item response theory (IRT) models as well as more contemporary extensions to the case of multilevel, multidimensional, and mixtures of discrete and continuous latent variables through the lens of discrete multivariate analysis. A general modeling framework is discussed, and the applications of this framework in diverse contexts are presented, including large-scale educational surveys, randomized efficacy studies, and diagnostic measurement. Other topics covered include parameter estimation and model fit evaluation. Both classical (numerical integration based) and more modern (stochastic) parameter estimation approaches are discussed. Similarly, limited information goodness-of-fit testing and posterior predictive model checking are reviewed and contrasted. The review concludes with a discussion of some emerging strands in IRT research such as response time modeling, crossed random effects models, and non-standard models for response processes.

1. INTRODUCTION

To the statistician, item response theory (IRT) typically represents a part of discrete multivariate analysis focused on modeling correlated multivariate response variables in a structured (often hierarchical) manner in terms of underlying latent variables of an either continuous (latent factors) or discrete (latent classes) nature. IRT is frequently used in educational or psychological testing or patient-reported health outcomes, in which data from questionnaires or standardized measurement instruments are modeled. In this review we keep these areas foremost in mind while trying to place the material in a broader statistical context.

In a more pragmatic sense, IRT encompasses a set of models along with the requisite computational and inferential tools that help researchers address technical issues such as item analysis, score reliability, and scale alignment that are related to the inherent fairness, quality, and validity questions (e.g., as defined by the joint standards on testing; AERA et al. 2014) associated with the development, administration, maintenance, interpretation, and use of tests. Particularly salient features of IRT models include the use of latent variables to represent the constructs being assessed, such as mastery, achievement, attitude, or severity of a disorder; the direct use of item-level (mainly categorical) data collected from the interactions of individual examinees, respondents, or patients with instances of assessment; and the reliance on conditional independence as the basic structure and recipe for model-building.

As Bock (1997) and Thissen & Steinberg (2009) noted, IRT has enjoyed continued development in the field of psychometrics, with many hundreds of papers and reports written over the last eight or nine decades, appearing frequently in outlets such as *Psychometrika*, the *Journal of Educational and Behavioral Statistics*, and the *Journal of Educational Measurement*. Book-length volumes on IRT, ranging from introductory (e.g., Hambleton et al. 1991) to more technical (e.g., Baker & Kim 2004), have been published. In many cases, the books have also witnessed revisions over the years to reflect growing developments in the research literature.

It is fair to say that IRT is now one of the central methodological pillars supporting many large and high-profile assessment programs around the globe, including the Organization for Economic Cooperation and Development (OECD) Program for International Student Assessment (PISA), the National Assessment of Educational Progress (NAEP, also known in the United States as the Nation's Report Card), the National Institutes of Health's Patient Reported Outcomes Measurement Information System initiative, and China's new National Assessment of Basic Education Quality. It has also become increasingly applied in academic research settings to enhance the precision and rigor of measurement in the social and behavioral sciences (e.g., Reise & Waller 2009).

Most of the applications of IRT methods to date are restricted to the case of unidimensional latent trait models for single-level data. That is, most of the time, researchers assume the presence of a single, continuous underlying latent variable that explains the covariation of the observed item responses. At the same time, the item response vectors are assumed to be independent across individuals. Hierarchically nested structures, such as students nested within classrooms or repeated measures nested within persons, are often ignored. Unfortunately, many constructs in the social and behavioral sciences are multifaceted (e.g., a chronic patient's quality of life), calling for multidimensional item response models. Multilevel item response models, on the other hand, provide mechanisms to represent the extra dependence introduced by nested data structures. Furthermore, a trait-like continuum representing individual differences is not the only kind of latent variable that is theoretically plausible and interpretable. For instance, models containing latent classes or mixtures of classification and continuous latent variables are needed in applications of IRT.

Because of IRT's disciplinary focus on the social sciences and its association with educational and psychological measurement in particular, the vocabulary and mathematical notation in primary sources can sometimes seem esoteric to the general applied statistics audience. This review attempts to avoid that by using more customary notation (e.g., $\boldsymbol{\theta}$ refers to parameters) and by beginning the discussion with a few motivating examples so that key concepts can be introduced.

2. SOME MOTIVATING EXAMPLES

2.1. A Beginning Example: Assessment Items in Education or Psychology

The item characteristic curve is one of the fundamental building blocks in the machinery of IRT. The item characteristic curve describes the conditional probabilities of observed item responses (e.g., correct versus incorrect response) given unobserved characteristics. In educational and psychological measurement, one of the most frequently encountered item characteristic curves is the three-parameter logistic (3PL) function used in this connection by Birnbaum (1968):

$$T_i(1|\eta) = \gamma_i + \frac{1-\gamma_i}{1+\exp\left[-(\alpha_i+\beta_i\eta)\right]}, \tag{1}$$

where the T_i notation pays homage to the earlier work on IRT by Paul Lazarsfeld (1950), who called the item characteristic curves trace lines because they trace the observable probability of a correct, or endorsement, response given unobserved variable η. The three parameters α_i (intercept), β_i (slope, or discrimination), and γ_i (pseudoguessing) are specific to item i, and as such they are referred to as item parameters, in contrast to those distributional parameters associated with specification of the prior density of the latent variable η (e.g., the mean and variance of η). If a random sampling interpretation is taken, the distribution of η can also be called the population density. The intercept is inversely associated with the degree of difficulty of the item. Because the rate of change in correct response probability moves faster for an item with a larger slope parameter, the slope is directly related to how well the item can differentiate among people having different degrees of proficiency. The pseudoguessing parameter reflects the nonzero probability of observing correct answers even for students with low proficiency (e.g., by randomly guessing). In psychological assessments, when IRT is used to model items about symptoms of a psychological disorder, it is also possible that some items will be endorsed even when the individual's overall symptom severity is low, potentially due to other unmodeled external causes (see, e.g., Reise & Waller 2009). Computationally and statistically efficient estimation of item and other distributional parameters as well as their standard errors is central to IRT modeling.

The 3PL function shown in Equation 1 has straightforward interpretations. For the purpose of illustration, consider the following example. A total of 5,750 elementary school students were given a set of state mathematics assessment items being field tested. There were 79 multiple-choice items scored dichotomously as correct (1) versus incorrect (0). **Figure 1*a*** shows the estimated 3PL item characteristic curve for item 75. The underlying latent variable could be understood as math proficiency. By convention in IRT, one usually fixes the location and scale of the latent variable by imposing a standard normal prior on η. With the method of maximum marginal likelihood (Bock & Aitkin 1981), the intercept is estimated to be -0.18 (SE $=$ 0.13), with a slope of 1.46 (SE $=$ 0.11), and the pseudoguessing probability is 0.28 (SE $=$ 0.03). As one can see, the curve takes on a sigmoidal shape, with a noticeable nonzero lower asymptote. Item 75 is not overly difficult as it has a middling intercept value. From experience, item 75 is also a reasonably discriminating item.

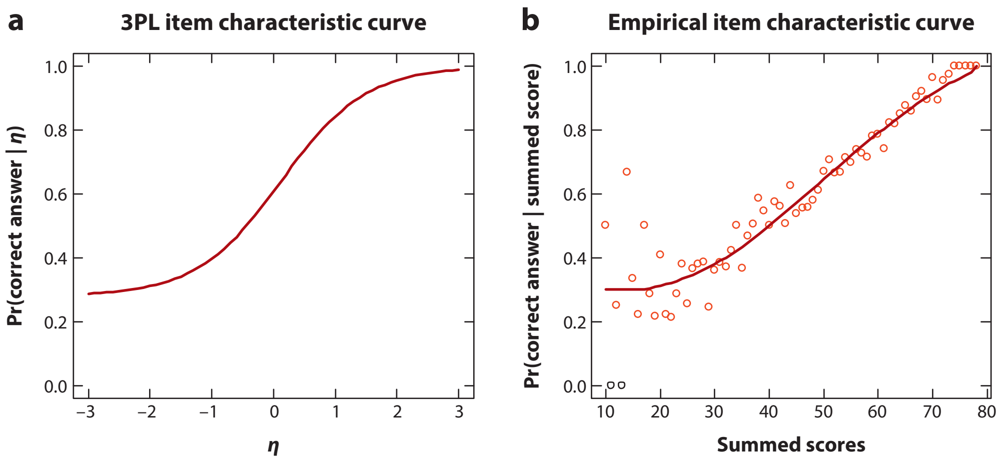

Figure 1

(*a*) Example three-parameter logistic (3PL) item characteristic curve with (*b*) its empirical counterpart. Panel *a* shows a classical 3PL item response function. The model can be justified on empirical grounds from the observed relationship between the proportion of correct answers and the summed scores (panel *b*).

The presence of substantially nonzero correct response probability toward the lower end of the summed score scale indicates the effect of guessing.

Simpler models for dichotomous responses do exist. For instance, the lower asymptote may be constrained to 0, which results in the two-parameter logistic (2PL) model. The item slopes may also be constrained to equality across items, resulting in the one-parameter logistic model, which is the identical twin of the Rasch model (Rasch 1961). One could also use altogether different link functions (e.g., the probit) to derive the more traditional normal ogive IRT models instead of their logistic counterparts. The motivation and derivation of these various models may be pursued either along the lines of the Rasch measurement tradition or as direct descendants and extensions of Thurstone's (1925) models (see, e.g., Thissen & Wainer 2001, Thissen & Steinberg 2009).

Yet, perhaps a more direct rationale for the use of these sigmoidal item response functions can be gleaned from an entirely empirical observation that should appeal to the applied statistician. **Figure 1*b*** shows an empirical item characteristic curve formed by tabulating the raw item response data. Setting item 75 aside, there are 78 other items. Students' observed total (summed) item scores may serve as a close proxy (assuming that the latent variable is unidimensional) for the unobserved proficiency variable η. The probabilities of responding correctly within each of the distinct summed score groups (ranging from 10 to 78 in this case), when plotted against the summed scores, follow a familiar S-shaped curve. For very low summed scores, the frequencies in those summed score groups are also small, resulting in more variability at the lower end. What is superimposed on the data points is the 3PL model–implied correct response probabilities for each of the summed score groups, based on the estimated item parameter values shown earlier. These probabilities are computed using a recursive algorithm first described by Lord & Wingersky (1984). Such plots and summaries are not only helpful as empirical motivations for IRT modeling, but they also form the basis of a number of model fit diagnostic indices (see Cai 2015). Extending this example, more complex models and applications are reviewed below, moving from dichotomous variables to polytomous variables, unidimensional to multidimensional, and single level to multilevel.

2.2. Large-Scale Educational Surveys

Large-scale educational surveys play an important role in research and policy discussions. In the United States, people carefully watch over long-term trajectories and changes in student achievement as measured by the NAEP, a congressionally mandated program. Across national boundaries, high-profile rankings based on the PISA, commissioned by the OECD, have sparked heated debates in the last decade about national educational competitiveness.

In order to support the NAEP, research conducted at the Educational Testing Service in the 1990s, led by pioneers such as Robert Mislevy, resulted in a general approach based on IRT and the plausible value (PV) methodology. The PV methodology is derived from Rubin's (Rubin 1987, Little & Rubin 2002) multiple imputation theory for analysis involving missing data. By NAEP standards, student achievement outcomes are considered latent variables and are hence unobserved or missing data. Student responses to assessment items and individual background information such as demographic variables and upper-level characteristics (e.g., at the school level) fill the role of observed data. To minimize respondent burden and cost and, more importantly, to provide efficient estimates of means, variance components, coefficients, and contrasts at levels of aggregation (e.g., districts, states, and countries, in the case of PISA) where principal policy interests lie, the design of large-scale educational assessments as exemplified by the NAEP also makes use of sampling of items so that each student receives only a small number of items, resulting in high fractions of missing data at the individual level. But estimates of the population parameters are biased when individual-level fractions of missing information are high (Mislevy 1991).

Rather than a single point estimate of student-level proficiency, multiple PVs are imputed from a posterior distribution that includes the student's item responses and background characteristics (Mislevy 1991, Mislevy et al. 1992). Central to the PV methods is a latent regression model that has components from both IRT and the general linear model.

Let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ be the matrix of background covariates, where $\boldsymbol{x}_j$ is the covariate vector for individual j and N is the total sample size. The background covariates may include information about demographics, schooling characteristics, motivation, resources, peer contexts, and so on. Let $\boldsymbol{y}_j = (\boldsymbol{y}_{j1}, \ldots, \boldsymbol{y}_{jD})$ be the individual assessment outcomes on D domains of interest (which are principally defined by policy interests, e.g., reading, math, and science for PISA), composed of item responses $\boldsymbol{y}_{jd}$ with typical element $\boldsymbol{y}_{ijd}$ denoting item i within domain d, and $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ collects all observed outcomes. We are interested in estimating the unobserved student achievement in D domains, and the vector of these latent variables for student j is represented by $\boldsymbol{\eta}_j = (\eta_{j1}, \ldots, \eta_{jD})$.

In modeling student responses, multiple categorical items are frequently encountered in addition to the dichotomously scored multiple-choice items. The multiple categorical responses may result from ordinal ratings of constructed responses such as essays or short answers. The generalized partial credit (GPC) model (Muraki 1992) is a widely used model for such item responses. Suppose item i has C_i categories, with item codes $c \in \{0, \ldots, C_i - 1\}$. A version of the GPC model may be written as

$$T_i(c \mid \eta) = \frac{\exp(c\beta_i\eta + \alpha_{ic})}{\sum_{m=0}^{C_i-1} \exp(m\beta_i\eta + \alpha_{im})}, \tag{2}$$

where the overall item slope β_i serves the same role as the discrimination parameter of a 3PL model. The category-specific intercept α_{ic} parameters are not identified without further constraints, typically with the first intercept fixed to 0 ($\alpha_{i0} = 0$). More broadly, the GPC model is one of a large number of models in the nominal categories model family (for details, see Thissen et al. 2010).

In the NAEP-conditioned latent regression model that produces the PVs, the latent proficiency vector $\boldsymbol{\eta}_j$ is assumed to be normally distributed with conditional mean vector $\boldsymbol{\Gamma x}_j$ and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Gamma}$ represents a matrix of unknown regression coefficients (Thomas & Gan 1997). The background covariates contribute to the explanation of the individual-specific outcomes by giving each student potentially different conditional means. Let $f(\boldsymbol{\eta}_j; \boldsymbol{\Gamma x}_j, \boldsymbol{\Sigma})$ represent a multivariate normal density function with mean $\boldsymbol{\Gamma x}_j$ and covariance matrix $\boldsymbol{\Sigma}$. Suppose that the items in each of the proficiency domains are independent once the latent proficiency η_{jd} is accounted for; this is known as the independent cluster factor pattern in the factor analysis literature. The posterior distribution of the latent proficiencies $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N)$ is then proportional to

$$p(\boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}) \propto \prod_{j=1}^{N} f(\boldsymbol{\eta}_j; \boldsymbol{\Gamma}\mathbf{x}_j, \boldsymbol{\Sigma}) \prod_{d=1}^{D} f_d(\mathbf{y}_{jd}|\eta_{jd}), \tag{3}$$

from which PVs of $\boldsymbol{\eta}$ can be drawn.

The current NAEP machinery is not without its own limitations. The current framework defines the proficiency domains a priori, and fails to account for any conditional dependence between items on different proficiency exams such as the effect of testlets (Cai 2010a). The current formulation also does not allow for investigators to explore alternative parameterizations of the latent proficiency space, such as combining items from different domains. Fully multidimensional IRT models with simultaneous estimation represent one avenue to resolve these difficulties.

2.3. Quantifying Effectiveness in Multisite Randomized Trials

Over the last decade or more, US educational research policy took a dramatic shift toward the use of randomized experiments to evaluate the effectiveness of intervention programs. In a review of trials funded by the US Department of Education, the majority were multisite cluster randomized experiments (Spybrook & Raudenbush 2009). In multisite cluster randomized trials, clusters (e.g., intact classrooms) within a site (e.g., a school) are randomly assigned to different conditions. Among its many benefits, the multisite trial essentially entails the use of mini-experiments at each site, which can be helpful in examining the generalizability of experimental results (Raudenbush & Liu 2000).

Multisite cluster randomized designs in the social and behavioral sciences typically include a test–retest administration of (almost always) identical assessments of the outcome construct to the same group of individuals at two time points, before and after the intervention. For example, the outcome could be a measure of math proficiency that the intervention is supposed to influence at a particular grade of interest.

In practice, the most common method for scoring outcome measures uses the summed score; that is, the total score from an assessment instrument is computed and used as observed outcome or predictor values. Researchers have also increasingly utilized IRT-based approaches (Curran et al. 2008) and estimated individual scaled scores such as expected a posteriori (EAP) estimates (Thissen & Wainer 2001). The IRT scaled scores are then treated as continuous outcome or predictor variables in subsequent data analyses.

The EAP method is not difficult to understand. Given estimated item parameters, the shape of the item characteristic curves is assumed predetermined. With the strong assumption of conditional independence of the item responses, one may write the joint probability of observing the individual vector of item responses $\boldsymbol{y}_j = (y_{ij}, \ldots, y_{nj})$ as

$$f(\boldsymbol{y}_j|\eta) = \prod_{i=1}^{n} f(y_{ij}|\eta), \tag{4}$$

where the conditional probability mass function of an individual item response is that of a multinomial with trial size 1 in C_i categories,

$$f(y_{ij}|\eta) = \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})}, \tag{5}$$

with the indicator function

$$\chi_c(y_{ij}) = \begin{cases} 1, & \text{if } c = y_{ij} \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Here, the assumption of conditional item independence is explicit and the posterior of the latent variable η is

$$f(\eta|\boldsymbol{y}_j) = \frac{\prod_{i=1}^{n} f(y_{ij}|\eta)\phi(\eta)}{\int \prod_{i=1}^{n} f(y_{ij}|\eta)\phi(\eta)d\eta}, \tag{7}$$

where $\phi(\eta)$ is the prior (population) distribution of the latent variable. Whereas $f(\eta|\boldsymbol{y}_j)$ must typically be approximated numerically (with quadrature), the computation of the posterior mean and variance is straightforward. Importantly, the EAP estimator shrinks individual scaled scores toward a common grand mean, as specified in the population distribution.

For a number of reasons, the existing IRT scaled score computation methods are problematic for randomized multisite intervention studies, despite being straightforward. First, the outcome constructs at each occasion are correlated due to the longitudinal design. The dependence of latent constructs pre- and posttreatment makes the choice of appropriate sample for item analysis ambiguous. Second, there is likely substantial item-level residual dependence because of repeated (pre-/post-) administration of the same items to the same individuals. If unattended to, the item-level residual dependence might lead to an overly optimistic estimate of statistical uncertainty. Third, the individuals are nested in sites, so there is likely extra correlation that should be modeled. Failure to account for such nesting may lead to the lack of congruence between statistical models used for measurement (single-level IRT models) and impact estimation (typically, multilevel hierarchical models). Last but not least, it is implausible to assume full exchangeability of individuals across treatment and control conditions. That is, the random variables associated with posttreatment individuals in the control condition cannot be exchanged without consequence to those in the treatment condition (see, e.g., Lindley & Smith 1972). What is needed in scoring is to assume conditional exchangeability, conditioning on the treatment assignment indicator, which is unfortunately something standard EAP calculations do not account for.

The last issue of exchangeability is particularly important, because failure to include key conditioning information (e.g., treatment assignment) in the measurement model can and will lead to inconsistencies in the subsequent inference about treatment effect. A more direct approach proposed by Cai et al. (2016), using a multilevel extension of Cai's (2010a) two-tier item factor analysis model, is illustrated below as a tool that systematically addresses these measurement issues in randomized evaluation studies of the efficacy of interventions.

2.4. Diagnostic Measurement

Diagnostic classification models (DCMs; see, e.g., Rupp et al. 2010) have received increasing attention within the field of educational and psychological measurement. In contrast to more traditional IRT models discussed thus far in which continuous latent variables are specified, DCMs presuppose the presence of discrete latent variables or attributes (e.g., mastery of skills or knowledge, presence of symptoms, possession of misconceptions) that explain the observed covariation of the item responses. For simplicity, consider a $Q \times 1$ vector of dichotomous (0–1) underlying latent variables, $\boldsymbol{\xi}$. The relationship between items and latent attributes is captured by the **Q**-matrix (Tatsuoka 1983), which is an $n \times Q$ matrix of fixed zeros and ones. The (i, q)-th entry in the **Q**-matrix takes on a value of 1 if item i measures attribute q.

Let $T_i(c\,|\boldsymbol{\xi})$ be the response function for item i in category $c \in \{0, \ldots, C_i - 1\}$. A useful model to consider for ordered polytomous item responses is the cumulative logit model based on Samejima's (1969) original graded response model. Let

$$T_i^+(c\,|\boldsymbol{\xi}) = \frac{1}{1 + \exp(-z_{ic})} \tag{8}$$

be the conditional cumulative response probability in categories c and above. Then the category response probability is the difference

$$T_i(c\,|\boldsymbol{\xi}) = T_i^+(c\,|\boldsymbol{\xi}) - T_i^+(c + 1|\boldsymbol{\xi}), \tag{9}$$

where $T_i^+(0|\boldsymbol{\xi}) = 1$ and $T_i^+(C_i|\boldsymbol{\xi}) = 0$ are the fixed boundary cases. In the case of DCMs, the linear predictor of the graded response model z_{ic} is

$$z_{ic} = \alpha_{ic} + \boldsymbol{\beta}_i' h_i(\mathbf{Q}, \boldsymbol{\xi}). \tag{10}$$

It is easy to see that the model contains $C_i - 1$ intercepts $(\alpha_{i1}, \ldots \alpha_{i,C_i-1})$. The $h_i(\mathbf{Q}, \boldsymbol{\xi})$ term in Equation 10 is a potentially vector-valued function that defines how the measured attributes combine to create the linear predictor portion of the item response model in Equation 8. As noted by Henson et al. (2009), constraints on the item parameters and choices of $h_i(\mathbf{Q}, \boldsymbol{\xi})$ yield several of the more commonly utilized diagnostic models (see also Rupp et al. 2010, Choi et al. 2013).

For instance, suppose that according to the **Q**-matrix, a mathematics test item i measures two particular algebra-related skills (attributes ξ_1 and ξ_2) and that successful solution of the item (so that the item score is $c = 1$) requires both attributes. The linear predictor may take the following deterministic-input noisy and (DINA) gate form (Junker & Sijtsma 2001), with an intercept and a single free parameter for the interaction term:

$$z_{i1} = \alpha_{i1} + \beta_i \xi_1 \xi_2. \tag{11}$$

In this case, $h_i(\mathbf{Q}, \boldsymbol{\xi}) = \xi_1\xi_2$ contains second-order interaction, and $\boldsymbol{\beta}_i = \beta_i$. Alternatively, perhaps the item only requires the mastery of either attribute 1 or attribute 2. Then the linear predictor of the IRT model may take the following deterministic-input noisy or gate form (Templin & Henson 2006), with the slope set equal between the main effects and the interaction in magnitude, albeit the interaction has a negative sign:

$$z_{i1} = \alpha_{i1} + \beta_i \xi_1 + \beta_i \xi_2 - \beta_i \xi_1 \xi_2. \tag{12}$$

In this case, $\boldsymbol{\beta}_i = (\beta_i, \beta_i, -\beta_i)'$ contains linear restrictions on the item parameters. Yet another possibility is that each attribute contributes to some increase in the log-odds of correct solution and that the magnitude of the increase due to one attribute does not necessarily require the mastery of the other. In that case, the linear predictor might contain only the main effect terms (Hartz 2002, von Davier 2005), similar to factor analysis:

$$z_{i1} = \alpha_i + \beta_i \xi_1 + \beta_i \xi_2. \tag{13}$$

DCMs can also be further integrated with item factor analysis models to provide even more flexible frameworks for data analysis, which is discussed next.

3. A MORE GENERAL FRAMEWORK

Given the foregoing discussion, a general modeling framework is presented below as a systematic synthesis of recent research on multidimensional, multilevel, and diagnostic classification IRT models. Consider a response in observed category c to item i. Suppose the level-one unit (respondent) j is nested within independent sampling unit (level-two unit) k, for example, a student nested within a classroom. The linear predictor can be written as

$$z_{ijkc} = \alpha_{ic} + \boldsymbol{\beta}'_{i1}\begin{pmatrix} h_{i1}(\mathbf{Q}_1, \boldsymbol{\xi}_{jk}) \\ \boldsymbol{\eta}_{jk} \end{pmatrix} + \boldsymbol{\beta}'_{i2}\begin{pmatrix} h_{i2}(\mathbf{Q}_2, \boldsymbol{\xi}_{k}) \\ \boldsymbol{\eta}_{k} \end{pmatrix}, \tag{14}$$

where $\boldsymbol{\beta}_{i1}$ and $\boldsymbol{\beta}_{i2}$ are the item slope parameters, with $\mathbf{Q}_1$ and $\mathbf{Q}_2$ the $\mathbf{Q}$-matrices at levels one and two, respectively. The subscripts on $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ indicate whether the latent variables vary over level-one units (jk) or over level-two units (k only). The dimensions of the slope vectors are dependent on both the number of continuous latent factors in $\boldsymbol{\eta}$ and the number of terms generated by the h functions at each level.

Substituting the linear predictor in Equation 14 into any one of the item response functions mentioned earlier (e.g., the 3PL), the conditional probability of observing response y_{ijk} remains that of a multinomial with trial size 1 in C_i categories (as in Equation 5):

$$f(y_{ijk}|\boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) = \prod_{c=0}^{C_i-1} T_i(c|\boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k)^{\chi_c(y_{ijk})}, \tag{15}$$

where the indicator function $\chi_c(y_{ijk})$ is similarly defined as in Equation 6. In order to build the model further, one needs to invoke the conditional independence assumption so that the conditional response pattern probability for response vector $\boldsymbol{y}_{jk} = (y_{1jk}, \ldots y_{njk})$ can be expressed as a product of individual item response probabilities:

$$f(\boldsymbol{y}_{jk}|\boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) = \prod_{i=1}^{n} f(y_{ijk}|\boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k). \tag{16}$$

Conditional independence is a strong assumption that deserves rigorous checking. In Section 5, model fit indices pertinent to examining conditional independence are reviewed. Regardless, in contrast to unidimensional IRT models reviewed above, the presence of both types of latent variables (continuous and discrete) at both levels help to better explain the covariation of item responses, making conditional independence more plausible in substantive measurement situations.

To complete the specification of the model, the relationships among discrete latent attributes are further modeled in a higher-order factor model (see, e.g., de la Torre & Douglas 2004). This amounts to treating the first-order latent variables ξ as though they were unobserved items and regressing them on other continuous (second-order) latent factors. For example, for dichotomous attributes, a multidimensional extension of the 2PL model (Reckase 2009) may be used to relate the latent attributes to the higher-order latent factors:

$$T_q(1|\boldsymbol{\zeta}_{jk}) = \frac{1}{1 + \exp[-(\tau_{q1} + \boldsymbol{\lambda}'_{q1}\boldsymbol{\zeta}_{jk})]}, \tag{17}$$

where τ_{q1} represents the latent attribute intercept and $\boldsymbol{\lambda}_{q1}$ is a vector of attribute slopes that describes how associated each of the higher-order factors are as related to the first-order attributes. Of course, Equation 17 is intended for latent variables at level one. For level-two attributes, we can define a similar expression:

$$T_q(1|\boldsymbol{\zeta}_{k}) = \frac{1}{1 + \exp[-(\tau_{q2} + \boldsymbol{\lambda}'_{q2}\boldsymbol{\zeta}_{k})]}, \tag{18}$$

regressing the level-two attributes on higher-order latent factors. Again, if we assume conditional independence of the latent attributes given $\boldsymbol{\zeta}$, we may write

$$\begin{aligned} f(\boldsymbol{\xi}_{jk}|\boldsymbol{\zeta}_{jk}) &= \prod_{q=1}^{Q_1} [T(1|\boldsymbol{\zeta}_{jk})]^{\xi_{jkq}} [1 - T(1|\boldsymbol{\zeta}_{jk})]^{1-\xi_{jkq}}, \\ f(\boldsymbol{\xi}_{k}|\boldsymbol{\zeta}_{k}) &= \prod_{q=1}^{Q_2} [T(1|\boldsymbol{\zeta}_{k})]^{\xi_{kq}} [1 - T(1|\boldsymbol{\zeta}_{k})]^{1-\xi_{kq}}, \end{aligned} \tag{19}$$

in order to fully represent the conditional attribute profile probabilities.

The key benefit to this higher-order factor model specification is that of simplicity. For q dichotomous attributes at either level, the unconstrained number of attribute profiles is equal to 2^q, with a huge number of parameters $(2^q - 1)$ even if the number of attributes is only reasonably large (e.g., a dozen). With the higher-order model and a single higher-order latent factor, the number of parameters is substantially reduced to $2q$.

The remaining task involves the conditioning of the η and ζ variables by regressing them on background covariates at levels one and two. Consider the latent regression of the η variables:

$$\begin{pmatrix} \boldsymbol{\eta}_{jk} \\ \boldsymbol{\eta}_{k} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ 0 & \boldsymbol{\Gamma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_{jk} \\ \boldsymbol{x}_{k} \end{pmatrix} + \begin{pmatrix} \epsilon_{jk} \\ \epsilon_{k} \end{pmatrix}, \tag{20}$$

where the lower left block of the regression coefficient matrix is constrained so that level-two latent variables are not directly regressed on level-one covariates. A similar equation can be specified for the higher-order factors ζ:

$$\begin{pmatrix} \boldsymbol{\zeta}_{jk} \\ \boldsymbol{\zeta}_{k} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ 0 & \boldsymbol{\Pi}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_{jk} \\ \boldsymbol{x}_{k} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\delta}_{jk} \\ \boldsymbol{\delta}_{k} \end{pmatrix}. \tag{21}$$

The equation disturbance terms have zero means and error covariance matrices that may contain estimable parameters subject to identification.

Putting various parts of the model together, and following standard modeling practices, the latent variables can be integrated out of the joint likelihood sequentially. Let $\boldsymbol{\theta}$ denote a $d \times 1$ vector that collects all free parameters in the model. These include parameters from all the items (α, β, and γ), parameters for the latent variables, the higher-order model (λ), and the regression coefficients and error covariance matrices. First, the level-one discrete latent attributes are integrated out:

$$f_\theta(\boldsymbol{y}_{jk}|\boldsymbol{\zeta}_{jk}, \boldsymbol{\xi}_{k}, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_{k}) = \int f_\theta(\boldsymbol{y}_{jk}|\boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_{k}, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_{k}) f_\theta(\boldsymbol{\xi}_{jk}|\boldsymbol{\zeta}_{jk}) d\boldsymbol{\xi}_{jk}. \tag{22}$$

Second, the level-one continuous latent variables are integrated out, bringing in the conditioning covariates as well:

$$f_\theta(\boldsymbol{y}_{jk}|\boldsymbol{\xi}_{k}, \boldsymbol{\eta}_{k}, \boldsymbol{x}_{jk}, \boldsymbol{x}_{k}) = \iint f_{\boldsymbol{\theta}}(\boldsymbol{y}_{jk}|\boldsymbol{\zeta}_{jk}, \boldsymbol{\xi}_{k}, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_{k}) f_\theta(\boldsymbol{\eta}_{jk}|\boldsymbol{x}_{jk}, \boldsymbol{x}_{k}) f_{\boldsymbol{\theta}}(\boldsymbol{\zeta}_{jk}|\boldsymbol{x}_{jk}, \boldsymbol{x}_{k}) d\boldsymbol{\eta}_{jk} d\boldsymbol{\zeta}_{jk}. \tag{23}$$

In the next step, one requires the assumption of conditional independence again, this time of the independence of N_k level-one response vectors nested within level-two unit k, conditionally on all the level-two latent variables and the background covariates:

$$f_{\boldsymbol{\theta}}\left(\{\boldsymbol{y}_{jk}\}_{j=1}^{N_k} | \{\boldsymbol{x}_{jk}\}_{j=1}^{N_k}, \boldsymbol{x}_{k}\right) = \iint \left[\int \prod_{j=1}^{N_k} f_{\boldsymbol{\theta}}(\boldsymbol{y}_{jk}|\boldsymbol{\xi}_{k}, \boldsymbol{\eta}_{k}, \boldsymbol{x}_{jk}, \boldsymbol{x}_{k}) f_{\boldsymbol{\theta}}(\boldsymbol{\xi}_{k}|\boldsymbol{\zeta}_{k}) d\boldsymbol{\xi}_{k} \right] f_{\boldsymbol{\theta}}(\boldsymbol{\zeta}_{k}|\boldsymbol{x}_{k}) f_{\boldsymbol{\theta}}(\boldsymbol{\eta}_{k}|\boldsymbol{x}_{k}) d\boldsymbol{\zeta}_{k} d\boldsymbol{\eta}_{k}, \tag{24}$$

where the parenthetical notation $\{\boldsymbol{y}_{jk}\}_{j=1}^{N_k}$ and $\{\boldsymbol{x}_{jk}\}_{j=1}^{N_k}$ simply collects together all the level-one outcome and covariate observations within level-two unit k.

Suffice it to say that Equation 24 contains the marginal likelihood function of all parameters of the model, if the observations are treated as fixed. Summing across the (assumed) independent level-two units, the overall marginal log-likelihood function is

$$\log L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) = \sum_{k=1}^{K} \log f_{\boldsymbol{\theta}}\left(\{\boldsymbol{y}_{jk}\}_{j=1}^{N_k} | \{\boldsymbol{x}_{jk}\}_{j=1}^{N_k}, \boldsymbol{x}_k \right). \tag{25}$$

Maximization of the log-likelihood typically requires numerical integration. A number of computational algorithms may be used, including the venerable expectation-minimization (EM) algorithm with fixed quadrature proposed by Bock & Aitkin (1981) and the more recent stochastic approximation-based Metropolis–Hastings Robbins–Monro (MH-RM; Cai 2010b,c) algorithm. A comparative review involving both EM and MH-RM algorithms is provided by Cai & Thissen (2014), and newer IRT software programs such as IRTPRO (Cai et al. 2011a) and flexMIRT® (Houts & Cai 2013) implement a large subset of the full features of the generalized modeling framework. Alternatively, one may also adopt fully Bayesian estimation for maximal flexibility in model specification and parameter estimation (see Edwards 2010 for a review in the context of IRT).

As mentioned earlier, the assumption of conditional independence is salient in the modeling framework described above. The degree to which subsequent model-based inferences are valid hinges on whether conditional independence is tenable. In practice, it is critical to check whether the latent variable dimensionality is correctly specified. In discussions of IRT model fit evaluation in Section 5.2, statistics that are sensitive to dimensionality misspecification are discussed.

4. ILLUSTRATIVE APPLICATIONS

The general model presented in Section 3 is not entirely identified without further constraints and specialization. For example, in typical applications, the distribution of latent variables and disturbance terms will contain some fixed means and covariance components in order to identify the location and scale of the unobserved variables in the model. Multidimensional IRT models also require sufficient constraints in order to avoid rotation and reflection indeterminacy. A set of applications is presented next, partly as continuations of the motivating examples from earlier discussions and partly to illustrate the broad applicability of the modeling framework reviewed here. They are not meant to be exhaustive.

4.1. Two-Tier Item Factor Model for Large-Scale Educational Assessments

Full-information item factor analysis (Bock et al. 1988) has become an important tool in large-scale educational assessment research. A special model, the item bifactor model (Gibbons & Hedeker 1992, Gibbons et al. 2007, Cai et al. 2011b), has also increasingly captured psychometricians' attention (Reise 2012). In a bifactor model, there is one primary factor, representing a target construct being measured, and there are S group–specific factors that are conditionally independent given the general factor. An item may load on at most one group-specific factor. The bifactor pattern is an example of a hierarchical solution (Holzinger & Swineford 1937, Schmid & Leiman 1957), dating back to earlier days in the factor analysis tradition.

Cai's (2010a) two-tier item factor model is a more general version of the bifactor model in that the number of general factors can be greater than one. In a two-tier model, two kinds of latent variables are considered, primary $\boldsymbol{\eta}_G$ and group-specific $\boldsymbol{\eta}_S$. In most cases, the group-specific factors are mutually orthogonal, and also orthogonal to the primary dimensions as an identification restriction (Rijmen 2009). In a two-tier model, an item may load on all primary factors, subject to model identification (such as rotational indeterminacy), and on at most one group-specific factor.

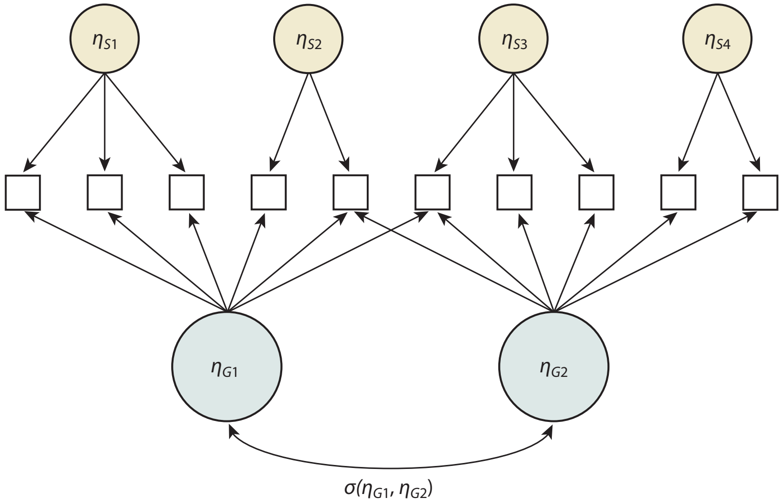

Figure 2

A two-tier model with two correlated primary factors and four group-specific factors. The items are shown as rectangles. The two primary factors are shown as the η_Gs below the items. The group-specific factors each influence a nonoverlapping set of items.

Figure 2 represents a hypothetical two-tier model with 10 items (the rectangles) that load on two correlated primary factors, as well as four group-specific factors that are mutually orthogonal. In terms of the factor pattern, the 10 by 6 factor pattern matrix corresponding to the model in **Figure 2** has the following form:

$$\begin{pmatrix} \beta & & \beta^* & & & \\ \beta & & \beta^* & & & \\ \beta & & \beta^* & & & \\ \beta & & & \beta^* & & \\ \beta & \beta & & \beta^* & & \\ \beta & \beta & & & \beta^* & \\ & \beta & & & \beta^* & \\ & \beta & & & \beta^* & \\ & \beta & & & & \beta^* \\ & \beta & & & & \beta^* \end{pmatrix},$$

where nonempty entries indicate free parameters, and β simply denotes a free slope parameter and does not imply equality constraints. The importance of the two-tier, or item bifactor, constraints is also a computational one. Maximum marginal likelihood estimation of the two-tier model requires at most one more fold of integration than the number of primary factors. Substantial computational time savings can be accomplished if software programs take advantage of this fact.

The two-tier model may also be useful for testlet-based educational assessments. Take the PISA reading literacy assessment as a case in point. The test design (see, e.g., Adams & Wu 2002) calls for several potentially correlated reading process dimensions (subscales) that are thought to contribute to the domain specification of reading literacy. These subscales are interpretation, reflection/evaluation, and retrieval of information. In addition, the format of the assessment utilizes passage-based reading tests with several questions following each prompt (forming into

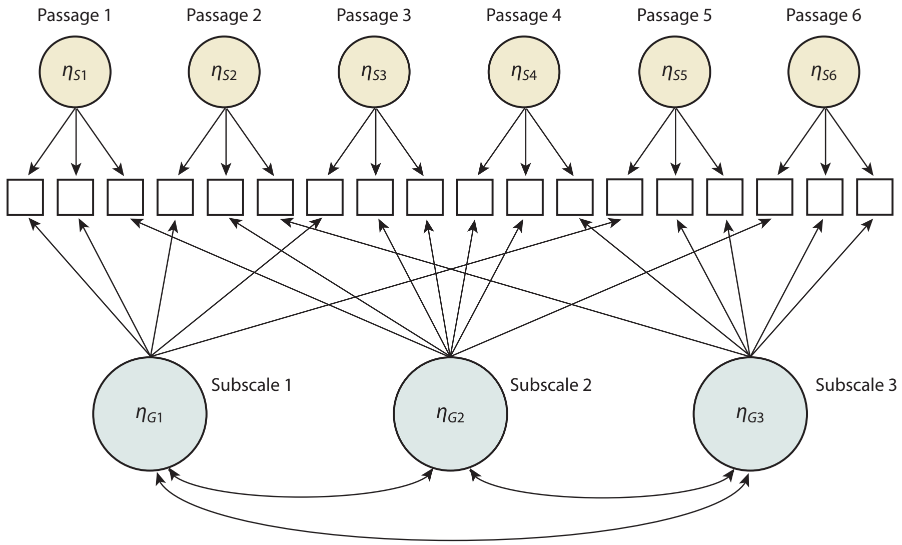

Figure 3

A two-tier model with potential application to testlet-based assessment. In the context of a typical passage-based reading assessment (e.g., the Program for International Student Assessment), the test may measure several related constructs (the primary factors), and at the same time one must account for the residual and unique influence of passages on item responses.

testlets). This design introduces extra dependence among the items in the same testlet because they all depend on the specific content of the same passage, which should be accounted for in IRT modeling.

Figure 3 shows a potential model for such an assessment. The three primary factors correspond to the three reading process subscales. Items in the same testlet can belong to different subscales, and the total number of group-specific factors is equal to six, one for each task. The two-tier model is accounting for multidimensionality/dependence within and/or between the correlated subscales. Furthermore, though not shown in the figure, it is possible to regress the reading process factors on background covariates from the PISA context questionnaire, for example, student academic interest and motivation factors. Harrell (2015) fitted such a model using NAEP data and the MH-RM algorithm.

Cai (2010a) fitted two-tier models to PISA reading and math assessment data simultaneously and not only found that the model offered improved fit, and could account for the substantial effect of residual item dependence within testlets, but also offered improved estimates of target dimensions of interest. This is because the two-tier item factor model can utilize the estimated factor correlations (often substantial across domains or subdomains) to borrow strength from other dimensions in the model to produce more precise marginal estimates of the latent variables. This is a familiar phenomenon in the context of shrinkage and subscore computations (Wainer et al. 2001, Haberman 2008).

4.2. Measurement and Efficacy of Intervention Studies

Presented here is an application of a multilevel two-tier item factor model to a large-scale randomized trial examining the instructional effects of serious video games, with over 1,500 students in 30 intervention and 29 comparison classrooms in 26 schools in 9 districts (Chung et al. 2014). The schools are the experimental sites, and intact classrooms within schools are randomly assigned. These serious learning games aim to improve students' knowledge of pre-algebra topics such as rational numbers and fractions. Students' mathematics learning outcomes were measured by items similar to prealgebra standardized assessment items (e.g., what might be included in state end-of-year tests), both before and after the experiment. Students in intervention classrooms played games on the topic of rational numbers and fractions, whereas those in the comparison classrooms played an alternative set of games on unrelated topics. This design is representative of a large class of evaluation designs used in federally funded education research. The hypothesis of the research team was that the students in intervention classrooms would do better overall than their peers in the control classrooms, when pretest differences were controlled for.

Without IRT, let Y_{ij} denote the observed summed score from student i in school j on the pre- and posttest assessments. The traditional hierarchical linear model (HLM; Raudenbush & Bryk 2002) is widely used to estimate the impact of the intervention. The observed pretest summed score is entered into the model as a covariate. The pretest scores are school-mean centered so the random coefficient β_{0j} could represent the mean posttest score for school j. The parameter of interest, β_{1j}, is the expected difference in posttest means between the two groups in school j, holding the pretest constant. The HLM may be compactly represented as

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} * Trt_{ij} + \beta_{2j} * \text{Pretest}_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \\ \beta_{0j} &= \gamma_{00} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}), \\ \beta_{1j} &= \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_{11}), \\ \beta_{2j} &= \gamma_{20} + u_{2j}, \quad u_{2j} \sim N(0, \tau_{22}). \end{aligned} \tag{26}$$

The fixed regression coefficients, β, and the variance components, τ, are estimated directly using maximum likelihood. The overall treatment impact estimate is focused on the regression parameter γ_{10}, which is estimated to be 1.12 (SE $=$ 0.33). The posttest standard deviation is 4.99, and one can convert the unstandardized regression coefficient into an estimated standardized effect size (Cohen's $d = 1.12/4.99 = 0.22$). Although it represents a statistically significant effect, the standardized effect size would be considered small to moderate in size. A concern the research team had was that the assessments were not long and might contain a substantial degree of measurement error. Because the assessment scores appear on both sides of the regression equation, the question is whether the results can be improved by applying more modern methods of data analysis such as IRT.

As mentioned earlier, classical summed (or standard IRT scaled) scores have a number of limitations when used in conjunction with cluster randomized multisite experimental studies. Cai (2010a) noted that in calibrating the IRT model with longitudinal item response data, even if the measurement instrument were unidimensional at each time point, the longitudinal item response data are inherently multidimensional. For designs with a pre- and posttest, at least two occasion-specific primary latent factors are needed for each of the treatment or control groups to model the initial status and potential gains. In addition, the responses to the same item in the pre- (time 1) and posttest (time 2) from the same individual may be residually dependent, even after controlling for the influence of the primary dimensions. Thus, item-specific residual dependence factors (doublets) are introduced to handle the potential residual dependence, and there are as many of them as the number of repeated items.

Multisite randomized experiments also generally consist of two or more groups—control and treatment in this case—within each site. Thus, it is necessary to specify four within-site primary dimensions to represent the pre- and posttest status of treatment and control groups. This is akin to conditioning the latent outcome variable on the treatment assignment indicator. Finally, from the study design, it is clear that multisite data have a nested structure, thus requiring both level-one and level-two (site) latent variables to model the site-level variation in initial status and change due to treatment receipt.

For a generic item i that appeared in both pre- and posttest, the linear predictor portions of the item response models could be written as follows:

$$\begin{aligned}
&\text{Pretest control:} \quad \beta_i[\eta_{1k} + \eta_{1jk}] + \beta_i^* \eta_{ijk}^*,\\
&\text{Posttest control:} \quad \beta_i[(\eta_{1k} + \eta_{1jk}) + (\eta_{2k} + \eta_{2jk})] + \beta_i^* \eta_{ijk}^*,\\
&\text{Pretest treatment:} \quad \beta_i[\eta_{1k} + \eta_{3jk}] + \beta_i^* \eta_{ijk}^*,\\
&\text{Posttest treatment:} \quad \beta_i[(\eta_{1k} + \eta_{3jk}) + (\eta_{2k} + \eta_{4jk})] + \beta_i^* \eta_{ijk}^*.
\end{aligned} \tag{27}$$

Each item has an overall discrimination parameter β_i, and the various latent variables contribute to the item response in a systematic manner. Specifically, two between-school latent variables, η_{1k} and η_{2k}, represent latent initial status and latent gain between pre- and posttest for school k, respectively. In addition, among the four within-school latent variables, the first two latent variables represent initial status (η_{1jk}) and latent gain (η_{2jk}) for student j in the control condition within school k, and the rest represent initial status (η_{3jk}) and latent gain (η_{4jk}) for student j in the treatment condition within school k. As such, the additional latent variables at posttest represent potential gains over the pretest level. Furthermore, the variation is decomposed at both pre- and posttest into between-school and within-school components. By allowing mean differences between η_2 and η_4 to be estimated, that is, the difference between latent changes between the treatment and control conditions within schools, potential effects of treatment on the gain in learning may be explicitly represented. The term η_{ijk}^* is an item-specific random effect that accounts for the residual dependence over repeated administrations of the same item in pre- and posttest.

This model is motivated by growth modeling developments as represented in Bock & Bargmann (1966), Embretson (1991), and McArdle (2009), among others. Upon estimating the item parameters, proper IRT scaled scores can then be computed for each of the latent variables as posterior means (EAP scores; Thissen & Wainer 2001) or as multiple imputations (PVs; see von Davier et al. 2009).

The same HLM seen in Equation 26 may be used to estimate the impact of the intervention. This time, the standard effect size in terms of Cohen's d is $0.24/0.42 = 0.57$. The substantial increase in estimated effectiveness stems from the fact that the latent variables serve to isolate the part of the posttest individual differences in learning outcomes that is sensitive to change, after controlling for prior knowledge. In contrast, the standard approach of using posttest scores (observed or scaled latent variable estimates) as an outcome conflates two sources of variation, namely, prior knowledge and potentially malleable difference. As a result, the latent gain score dimensions from the multilevel two-tier model are more sensitive measures for targeted interventions than the observed outcome scores (for a similar phenomenon, see Gibbons et al. 2008, p. 365).

Multilevel IRT models may also be used in observational studies. For instance, Yang & Cai (2014) examined the application of multilevel IRT models to the study of contextual effects. The estimation of these multilevel IRT models are nontrivial, and stochastic optimization algorithms are often necessary.

4.3. Improving Diagnostic Measurement

This example builds on prior work by Lee et al. (2011), who analyzed data from booklets 4 and 5 from the 2007 Trends in Mathematics and Science Study fourth-grade mathematics test. They had 25 test items reviewed and rated according to the specific testing objectives described in the assessment framework. Of these items, 15 unique testing objectives were identified. Accordingly,

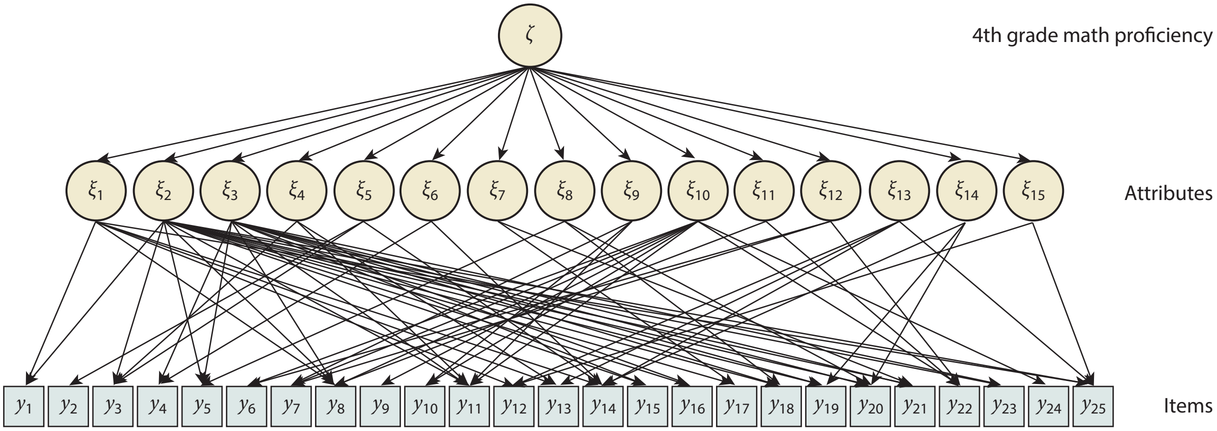

Figure 4

Path diagram of original Q-matrix and higher-order model structure. The 25 observed items are shown as the rectangles. They serve to measure 15 discrete attributes. The associations among the attributes are further modeled with a higher-order latent variable, ξ.

they set up a **Q**-matrix with 25 rows and 15 columns. The attributes represent the skills required in each of the testing objectives.

Lee et al. (2011) specified a conjunctive DINA model for the items in their analysis, which meant that skills in all testing objectives measured by an item had to be mastered for a correct solution. A higher-order DINA model (de la Torre & Douglas 2004) using the reported **Q**-matrix was specified and fitted to a sample of 564 US students. **Figure 4** presents a path diagram of this higher-order DCM. The latent attribute variables are shown as circles with a line across to indicate that they are discrete. The attribute variables are regressed further on a single higher-order factor that may be thought of as math proficiency. The item-to-attribute relationship is complex, with most items associated with a number of attributes. This is fairly typical of DCMs because the attributes represent narrow skills, in contrast to item factor analysis models that contain factors representing broad traits or abilities.

The items considered in this study have all since been made public (Foy & Olson 2009) and were thus available for review. Upon fitting the model, it was clear that the model failed to completely account for the dependence between items 18 and 19. Reviewing the content of the items showed that the two items are administered within a testlet (M031242A/B/C), and it was evident upon further inspection that a correct response to item 18 (M031242A) could greatly help in answering item 19. This observation may be the reason why the initial DCM could not explain the dependence between these two items. Adding another latent factor to explain the residual dependence of the item doublet results in the addition of a single parameter. This is akin to adding a random intercept as in Wainer and colleagues' (2007) testlet response theory model. A likelihood ratio test between the initial and the modified model showed that the addition of the extra parameter led to significant model fit improvement: $\Delta\chi^2(1) = 53$, $p < 0.001$.

5. ESTIMATION AND MODEL FIT

5.1. IRT Parameter Estimation

As Cai & Thissen (2014) mentioned in their recent review on IRT parameter estimation methods, a central idea that runs through parameter estimation in IRT is the missing data formulation, wherein the latent variables are regarded as missing observations that should be filled in. Once they are properly filled in, the remainder of parameter estimation becomes straightforward because one can take advantage of the conditional independence structure inherent in the model.

This can be illustrated with an example. For simplicity, consider a unidimensional IRT model. Given the foregoing discussions, the marginal probability associated with response pattern $\boldsymbol{y}_j = (y_{1j}, \ldots y_{nj})$ is

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}_j) = \int \prod_{i=1}^{n} \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})} f(\eta) d\eta = \int f_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \eta) d\eta. \tag{28}$$

If the item responses $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ are regarded as fixed once they are observed, the marginal likelihood function for all the item parameters in $\boldsymbol{\theta}$ can be expressed as

$$L(\theta|\mathbf{Y}) = \prod_{j=1}^{N} \int \prod_{i=1}^{n} \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})} f(\eta) d\eta. \tag{29}$$

Because the marginal likelihood $L(\boldsymbol{\theta}|\mathbf{Y})$ does not depend on the unobserved η values, it may be referred to as the observed data likelihood. The observed data likelihood can be approximated via numerical quadrature and directly optimized using Newton-type algorithms (Moustaki & Knott 2014) or the EM algorithm (Bock & Aitkin 1981).

Implicit in the "observed data" terminology is a realization that the latent variables are the missing data. Thus, the so-called complete data likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\eta}) = \left[\prod_{j=1}^{N} f(\eta_j)\right] \left[\prod_{j=1}^{N} \prod_{i=1}^{n} \prod_{c=0}^{C_i-1} T_i(c|\eta)^{\chi_c(y_{ij})}\right], \tag{30}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)$ collects all the latent variable scores. Equation 30 shows that the complete data likelihood takes a factored form because of the assumption of item conditional independence and the assumed independence of individuals. It is also easy to see that the complete data likelihood is proportional to the posterior distribution of $\boldsymbol{\eta}$, $p_{\boldsymbol{\theta}}(\boldsymbol{\eta}|\mathbf{Y})$. With the help of this posterior distribution, one may verify that given item parameter values, the following equality holds [it is known as Fisher's (1925) identity] assuming mild regularity conditions:

$$\frac{\partial \log f_{\boldsymbol{\theta}}(\boldsymbol{y}_j)}{\partial \boldsymbol{\theta}} = \int \frac{\partial \log f_{\boldsymbol{\theta}}(\boldsymbol{y}_j, \eta)}{\partial \boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\eta|\boldsymbol{y}_j) d\eta. \tag{31}$$

In other words, the gradient of the observed data likelihood is equal to the expectation of the complete data likelihood over the posterior distribution of the latent variables given the observed variables. This insight suggests that instead of maximizing the observed data likelihood, one could iteratively maximize the conditional expected complete data likelihood, eventually ending at the same solution.

Bock & Aitkin's (1981) solution can be considered a direct application of Fisher's identity. They noted that the marginal probability can be approximated by replacing the integral with a summation over a set of quadrature points spread over $\boldsymbol{\eta}$:

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}_j) \approx \bar{P}_j = \sum_{s=1}^{S} \prod_{i=1}^{n} \prod_{c=0}^{C_i-1} T_i(c|A_s)^{\chi_c(y_{ij})} W_s, \tag{32}$$

where A_s is a quadrature point and W_s is the corresponding weight (e.g., from Gauss–Hermite integration formulas). Bock & Aitkin (1981) also saw that the height of the posterior distribution at quadrature point A_s could be approximated:

$$p_{\boldsymbol{\theta}}(A_s|\boldsymbol{y}_j) \approx \frac{\prod_{i=1}^{n}\prod_{c=0}^{C_i-1} T_i(c|A_s)^{\chi_c(y_{ij})} W_s}{\bar{P}_j}. \tag{33}$$

Ignoring constants involving the prior distribution $f(\eta_j)$ from Equation 30, the complete data log-likelihood for the item parameters can be written as:

$$\log L(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\eta}) \propto \sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{c=0}^{C_i-1} \chi_c(y_{ij}) \log T_i(c|\eta). \tag{34}$$

Thus, the conditional expected complete data log-likelihood given provisional item parameter values $\boldsymbol{\theta}^*$ can be approximated as follows

$$Q(\boldsymbol{\theta}|\mathbf{Y};\boldsymbol{\theta}^*) \approx \sum_{j=1}^{N}\sum_{s=1}^{S}\sum_{i=1}^{n}\sum_{c=0}^{C_i-1} \chi_c(y_{ij}) \log T_i(c|A_s) p_{\boldsymbol{\theta}^*}(A_s|\boldsymbol{y}_j). \tag{35}$$

A third important insight from Bock & Aitkin (1981) is that by interchanging the order of summation, they realized that the posterior probabilities can be accumulated over individuals first:

$$Q(\boldsymbol{\theta}|\mathbf{Y};\boldsymbol{\theta}^*) = \sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{c=0}^{C_i-1} r_{isc} \log T_i(c|A_s), \tag{36}$$

where

$$r_{isc} = \sum_{j=1}^{N} \chi_c(y_{ij}) p_{\boldsymbol{\theta}^*}(A_s|\boldsymbol{y}_j)$$

represents the conditional expected probabilities of individuals in category *c* for item *i* at quadrature point A_s. As one can see, the conditional expected complete data model in Equation 36 is far more tractable than the observed data model in Equation 29 because the complete data model reduces to a series of independent logit analyses with A_s as the predictor values and r_{isc} as the outcome weights. In the E-step, the conditional expected probabilities are computed. In the M-step, the conditional expected complete data log-likelihood is maximized. The two steps are repeated until convergence.

Although the Bock–Aitkin EM algorithm elegantly handles unidimensional and sufficiently low-dimensional IRT parameter estimation, it does not generalize to the case of more complex IRT models such as the kind discussed in Section 3, because even with a moderate number of latent dimensions, the exponentially increasing size of the quadrature grid presents an insurmountable computational burden so far. Adaptive quadrature (see, e.g., Schilling & Bock 2005) has been employed to limited but not complete success. Wirth & Edwards (2007) referred to the problem as the "challenge of dimensionality."

Cai (2010b,c) noted that a solution already resides in Fisher's identity. If one could draw multiple random imputations of η from its posterior predictive distribution $p_{\boldsymbol{\theta}^*}(\eta|\boldsymbol{y}_j)$, with provisional item parameter estimates $\boldsymbol{\theta}^*$, the right-hand side of Equation 31 can be approximated by Monte Carlo, which is the same as the left-hand side over repeated sampling. In other words, the Monte Carlo average of complete data log-likelihood gradients gives the same likelihood ascent direction as the observed data log-likelihood gradient vector! With the widespread availability of Markov chain Monte Carlo (MCMC) methods (e.g., the Metropolis–Hastings methods; Metropolis et al. 1953, Hastings 1970), the imputations can be produced rather easily because the complete data likelihood is proportional to the posterior distribution of $\boldsymbol{\eta}$.

One problem still remains: The Monte Carlo approximation contains error, and unless the Monte Carlo size increases, the random sampling error obscures the true direction of likelihood ascent. This is a known issue in the context of Monte Carlo EM (Booth & Hobert 1999). Cai (2010b) noted that instead of trying to contain the Monte Carlo noise as a nuisance, it may in fact be employed productively. By drawing an analogy to Robbins & Monro's (1951) classical stochastic approximation (SA) algorithm, the Monte Carlo noise can be regarded as the stochastic excitations that drive an underlying dynamical system with the maximum marginal likelihood solution being a stability point. The noise can be filtered out step-by-step with the use of a sequence of slowly decaying gain constants. This results in the MH-RM algorithm that borrows elements of MCMC and combines them with SA. The MH-RM algorithm can be used to estimate all of the models described here, and because it eschews quadrature, it is computationally manageable even for high-dimensional IRT models.

5.2. IRT Model Fit Evaluation

Upon fitting the model, the issue of model fit evaluation becomes immediately relevant. Given that the models discussed here are highly parameterized and contain a large number of latent variables, and making rather strict assumptions related to distributional form and conditional independence, it is critical that overall model fit tests and diagnostic indices be developed. Unfortunately, this is an area wherein IRT modeling has historically lagged significantly behind other latent variable modeling frameworks (e.g., the structural equation model). In recent years, however, two strands of model fit evaluation research have emerged within IRT, providing promising tools that are being increasingly utilized in the data analytic practice.

The first strand involves limited-information goodness-of-fit testing (see Maydeu-Olivares & Joe 2005, Cai et al. 2006, Maydeu-Olivares 2013 and the references therein). The IRT model is built on an underlying multiway contingency table made up of the full item × item × item ... cross-classifications. For dichotomously scored items, the size of this contingency table grows exponentially as the number of items increases (2^n). With a modest number of items, say, 20, the total number of response pattern probabilities in the underlying general multinomial easily surpasses the sample size. Therefore, classical likelihood ratio or Pearson chi-square statistics no longer have their purported asymptotic chi-square null distributions as a result of sparseness (Bartholomew & Tzamourani 1999), though conditional inferences based on these statistics remain useful (McCullagh 1986). A more recent solution advocated by IRT researchers in recent years is to examine the lower-order marginal tables of the underlying multinomial because the cells in these lower-order (e.g., first- and second-order) margins are much better filled. Wald-type chi-squares can then be constructed (see, e.g., Cai & Hansen 2013). Despite an apparent loss of information (because the higher-order moments are ignored), the limited-information fit statistics do have much better calibration than the classical (full-information) counterparts and can also be more powerful against important alternatives that impact the lower-order marginal probabilities such as dimensionality misspecifications. One advantage of the limited-information model fit tests is that the test statistics can be easily decomposed into univariate and bivariate residuals (see, e.g., Liu & Maydeu-Olivares 2013, 2014) so that the sources of model misfit can be examined.

The second strand involves posterior predictive model checking (PPMC; Hoijtink & Molenaar 1997, Janssen et al. 2000, Glas & Meijer 2003, Sinharay et al. 2006, Levy et al. 2009, Levy & Svetina 2011). In PPMC, instead of relying on classical asymptotic arguments and chi-square indices, the researchers took the alternative stance that model fit evaluation can be determined on the basis of comparisons of realized versus observed discrepancies from repeated simulations of plausible (future) data sets under the fitted model. This inherently Bayesian concept (Rubin 1981, 1984) has penetrated IRT such that it has already been applied to the assessment of item-level model fit, person fit, and dimensionality. An interesting connection between the PPMC literature and the more classical model evaluation literature is in the choice of discrepancy measures. In PPMC, choices of discrepancy measures lead to differentially diagnostic indices sensitive against different forms of misspecification. Some classically derived indices such as Orlando & Thissen's (2000) item fit indices based on summed scores have been found successful as discrepancy measures.

6. DISCUSSION

After decades of development, IRT has become a staple in applied statistics in the social and behavioral sciences. This is further exemplified by the recent emergence of dedicated IRT procedures in large statistical software packages such as Stata® and SAS®. A plethora of R packages also exist (e.g., mirt; Chalmers 2012) that often offer enhanced features found in more recent research literature. The application of IRT has also expanded quite considerably from its original base in educational measurement. The diversification adds value to IRT, but the number of IRT models available limits coverage due to length constraints. Several new classes of IRT models are discussed below.

6.1. Response Time Models

With the proliferation of computer-based data collection, individual response time data are increasingly collected along with the responses themselves. Although the item-specific response time outcomes are no longer discrete data, response time data can be integrated into the IRT modeling framework and jointly modeled with item responses. Response time models of log-normal type (see, e.g., van der Linden 2006) have been developed to account for the time-intensity of items and individuals' speed. Response time models have also been used in IRT-based computerized adaptive testing (van der Linden 2008).

6.2. Crossed Random Effects

The IRT models discussed here are nested random effects models. There is an altogether different class of IRT models that regards the items as crossed with individuals (De Boeck 2008). In these models, item parameters are no longer fixed but are random variables themselves. Such a conceptual shift leads to an interesting generalizability potential with IRT models (Briggs & Wilson 2007). One significant application of crossed random effect IRT models is in the area of automatic item generation (Embretson 1999, Janssen et al. 2000, Glas & van der Linden 2003, Sinharay et al. 2003).

6.3. Special Response Processes

The IRT models discussed in this review contain relatively complex latent structural models but simple logit-based link functions. Models for special response processes exist. First, semiparametric or nonparametric item response functions have recently been considered (Liang 2007, Miyazaki & Hoshino 2009, Falk & Cai 2016) as a compromise between older approaches such as kernel smoothing (Ramsay 1991) and even older and more restrictive parametric approaches. There are also IRT models with semiparametric latent distributions (e.g., Woods & Thissen 2006, Monroe & Cai 2014). Second, there is another class of models that generalize the assumption of monotonic item response functions (dominance response process), the so-called ideal point (unfolding) models (e.g., Andrich 1996). In unfolding models, item response functions have peaks, reflecting ideal points of response along the latent dimension. There have also been some interesting developments in multidimensional unfolding models (e.g., Maydeu-Olivares et al. 2006). Third, the IRT models discussed thus far are not appropriate for forced-choice questionnaires. With forced-choice items, respondents must choose between two or more items presented at the same time. This introduces opportunities for improved measurement but also induces complex dependencies in the data. Brown (2016) offers a general framework for analyzing data arising from forced-choice items. Finally, in assessment contexts other than educational testing, individuals can and often do exhibit response styles, such as middle response, socially desirable response, and extreme response. Recently, formal models have been developed in IRT to handle the extra complexities induced by individual response styles (Böckenholt 2012, 2014; Thissen-Roe & Thissen 2013; Khorramdel & von Davier 2014; Plieninger & Meiser 2014; Falk & Cai 2016) with emerging utility for nontraditional applications of IRT.

Because of IRT's long history, the purpose of this review is a rather modest one. What has been offered is a general statistical framework based on recent technical developments in multidimensional and multilevel IRT modeling, along with a conceptual map that tries to unite several substrands of research within IRT. It is fair to conclude that the statistical field of IRT modeling will continue to enjoy active development in the coming decades as the cost of data collection drops further in the social and behavioral sciences and that technology enhances the role of evidence-based assessment and evaluation in decision making.

DISCLOSURE STATEMENT

Li Cai was a developer of two software packages referenced in this review (IRTPRO and flexMIRT). The remaining authors (Choi, Hansen, and Harrell) are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors' research is supported in part by a grant from the Institute of Education Sciences (R305D140046). The views expressed here belong to the authors and do not reflect the views or policies of the funding agency. The authors wish to thank Dr. Wenjing Huang for helpful comments and edits.

LITERATURE CITED

Adams RJ, Wu ML, eds. 2002. *PISA 2000 technical report.* Tech. Rep., Organ. Econ. Coop. Dev., Paris

AERA (Am. Educ. Res. Assoc.), APA (Am. Psychol. Assoc.), NCME (Natl. Counc. Meas. Educ.). 2014. *Standards for Educational and Psychological Testing*. Washington, DC: Am. Educ. Res. Assoc.

Andrich D. 1996. A hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Likert methodologies. *Br. J. Math. Stat. Psychol.* 49:347–65

Baker FB, Kim S-H. 2004. *Item Response Theory: Parameter Estimation Techniques*. New York: Dekker

Bartholomew, DJ, Tzamourani P. 1999. The goodness-of-fit of latent trait models in attitude measurement. *Sociol. Res.* 27:525–46

Birnbaum A. 1968. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*, ed. FM Lord, MR Novick, pp. 397–479. Reading, MA: Addison-Wesley

Bock RD. 1997. A brief history of item theory response. *Educ. Meas. Issues Pract.* 16:21–33

Bock RD, Aitkin M. 1981. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46:443–59

Bock RD, Bargmann RE. 1966. Analysis of covariance structures. *Psychometrika* 31:507–33

Bock RD, Gibbons RD, Muraki E. 1988. Full-information item factor analysis. *Appl. Psychol. Meas.* 12:261–80

Böckenholt U. 2012. Modeling multiple response processes in judgment and choice. *Psychol. Methods* 17:665–78

Böckenholt U. 2014. Modeling motivated misreports to sensitive survey questions. *Psychometrika* 79:515–37

Booth JG, Hobert JP. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B* 61:265–85

Briggs DC, Wilson M. 2007. Generalizability in item response modeling. *J. Educ. Meas.* 44(2):131–55

Brown A. 2016. Item response models for forced-choice questionnaires: a common framework. *Psychometrika* 81:135–60

Cai L. 2010a. A two-tier full-information item factor analysis model with applications. *Psychometrika* 75:581–612

Cai L. 2010b. High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika* 75:33–57

Cai L. 2010c. Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *J. Educ. Behav. Stat.* 35:307–35

Cai L. 2015. Lord–Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika* 80:535–59

Cai L, Choi K, Kuhfeld M. 2016. On the role of multilevel item response models in multi-site evaluation studies for serious games. In *Issues Regarding the Use of Games and Simulations for Teaching and Assessment*, ed. HF O'Neil, EL Baker, R Perez. New York: Taylor & Francis. In press

Cai L, Hansen M. 2013. Limited-information goodness-of-fit testing of hierarchical item factor models. *Br. J. Math. Stat. Psychol.* 66:245–76

Cai L, Maydeu-Olivares A, Coffman DL, Thissen D. 2006. Limited information goodness-of-fit testing of item response theory models for sparse 2^P tables. *Br. J. Math. Stat. Psychol.* 59:173–94

Cai L, Thissen D. 2014. Modern approaches to parameter estimation in item response theory. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*, ed. SP Reise, D Revicki, pp. 41–59. New York: Taylor & Francis

Cai L, Thissen D, du Toit SHC. 2011a. IRTPRO: flexible, multidimensional, multiple categorical IRT modeling. *Computer Software*. Lincolnwood, IL: Sci. Softw. Int.

Cai L, Yang J, Hansen M. 2011b. Generalized full-information item bifactor analysis. *Psychol. Methods* 16:221–48

Chalmers RP. 2012. mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48:1–29

Choi H-J, Rupp AA, Pan M. 2013. Standardized diagnostic assessment design and analysis: key ideas from modern measurement theory. In *Self-Directed Learning Oriented Assessment in the Asia-Pacific*, ed. R Maclean, pp. 61–85. New York: Springer

Chung GKWK, Choi K, Baker EL, Cai L. 2014. *The effects of math video games on learning: a randomized evaluation study with innovative impact estimation techniques*. CRESST Rep. 841, Natl. Cent. Res. Eval., Stand., Stud. Test., Univ. Calif., Los Angeles

Curran P, Hussong A, Cai L, Huang W, Chassin L, et al. 2008. Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Dev. Psychol.* 44(2):365–80

De Boeck P. 2008. Random item IRT models. *Psychometrika* 73:533–59

de la Torre J, Douglas JA. 2004. Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69:333–53

Edwards MC. 2010. A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika* 75:474–97

Embretson S. 1991. A multidimensional latent trait model for measuring learning and change. *Psychometrika* 56:495–515

Embretson SE. 1999. Generating items during testing: psychometric issues and models. *Psychometrika* 64:407–33

Falk CF, Cai L. 2016. Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*. In press. doi: 10.1007/s11336-014-9428-7

Fisher RA. 1925. Theory of statistical estimation. *Proc. Camb. Philos. Soc.* 22:700–25

Foy P, Olson JF. 2009. *TIMSS 2007 user Guide for the International Database*. Chestnut Hill, MA: TIMSS/PIRLS Int. Study Cent., Boston Coll.

Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, et al. 2007. Full-information item bifactor analysis of graded response data. *Appl. Psychol. Meas.* 31:4–19

Gibbons RD, Hedeker DR. 1992. Full-information item bi-factor analysis. *Psychometrika* 57(3):423–36

Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagiolini A, et al. 2008. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr. Serv.* 59:361–68

Glas CAW, Meijer RR. 2003. A Bayesian approach to person fit analysis in item response theory models. *Appl. Psychol. Meas.* 27:217–33

Glas CAW, van der Linden WJ. 2003. Computerized adaptive testing with item cloning. *Appl. Psychol. Meas.* 27:247–61

Haberman SJ. 2008. When can subscores have value? *J. Educ. Behav. Stat.* 33:204–29

Hambleton RK, Swaminathan H, Rogers HJ. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage

Harrell LA. 2015. *Analysis strategies for planned missing data in health sciences and education research*. PhD Thesis, Dep. Biostat., Univ. Calif., Los Angeles

Hartz SM. 2002. *A Bayesian framework for the unified model for assessing cognitive abilities blending theory with practicality*. PhD Thesis, Univ. Ill., Urbana-Champaign

Hastings WK. 1970. Monte Carlo simulation methods using Markov chains and their applications. *Biometrika* 57:97–109

Henson R, Templin JL, Willse JT. 2009. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74:191–210

Hoijtink H, Molenaar IW. 1997. A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika* 62:171–80

Holzinger KJ, Swineford F. 1937. The bi-factor method. *Psychometrika* 2:41–54

Houts CR, Cai L. 2013. *flexMIRT user's manual version 2.0: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychom. Group

Janssen R, Tuerlinckx F, Meulders M, De Boeck P. 2000. A hierarchical IRT model for criterion-referenced measurement. *J. Educ. Behav. Stat.* 25:285–306

Junker B, Sijtsma K. 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25:258–72

Khorramdel L, von Davier M. 2014. Measuring response styles across the big five: a multiscale extension of an approach using multinomial processing trees. *Multivar. Behav. Res.* 49(2):161–77

Lazarsfeld PF. 1950. The logical and mathematical foundations of latent structure analysis. In *Measurement and Prediction*, Vol. 4: *Studies in Social Psychology in World War II*, ed. SA Stouffer SA, L Buttman, EA Suchman, PF Lazarsfeld, SA Star, JA Clausen, pp. 362–412. Princeton, NJ: Princeton Univ. Press

Lee YS, Park YS, Taylan D. 2011. A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *Int. J. Test.* 11:144–77

Levy R, Mislevy RJ, Sinharay S. 2009. Posterior predictive model checking for multidimensionality in item response theory. *Appl. Psychol. Meas.* 33:519–37

Levy R, Svetina D. 2011. A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *Br. J. Math. Stat. Psychol.* 64:208–32

Liang L. 2007. *A semi-parametric approach to estimating item response functions.* PhD Thesis, Dep. Psychol., Ohio State Univ., Columbus, OH

Lindley DV, Smith AFM. 1972. Bayes estimates for the linear model (with discussion). *J. R. Stat. Soc. Ser. B* 34:1–41

Little RJA, Rubin DB. 2002. *Statistical Analysis with Missing Data*. New York: Wiley. 2nd ed.

Liu Y, Maydeu-Olivares A. 2013. Local dependence diagnostics in IRT modeling of binary data. *Educ. Psychol. Meas.* 73(2):254–74

Liu Y, Maydeu-Olivares A. 2014. Identifying the source of misfit in item response theory models. *Multivar. Behav. Res.* 49:354–71

Lord FM, Wingersky MS. 1984. Comparison of IRT true-score and equipercentile observed-score "equatings." *Appl. Psychol. Meas.* 8: 453–61

Maydeu-Olivares A. 2013. Goodness-of-fit assessment of item response theory models. *Meas. Interdiscip. Res. Perspect.* 11:71–101

Maydeu-Olivares A, Hernández A, McDonald RP. 2006. A multidimensional ideal point IRT model for binary data. *Multivar. Behav. Res.* 44:445–72

Maydeu-Olivares A, Joe H. 2005. Limited and full information estimation and goodness-of-fit testing in 2^n contingency tables: a unified framework. *J. Am. Stat. Assoc.* 100:1009–20

McArdle JJ. 2009. Latent variable modeling of difference and changes with longitudinal data. *Annu. Rev. Psychol.* 60:577–605

McCullagh P. 1986. The conditional distribution of goodness-of-fit statistics for discrete data. *J. Am. Stat. Assoc.* 81:104–7

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equations of state space calculations by fast computing machines. *J. Chem. Phys.* 21:1087–92

Mislevy RJ. 1991. Randomization-based inference about latent variables from complex samples. *Psychometrika* 56:177–96

Mislevy RJ, Johnson EG, Muraki E. 1992. Scaling procedures in NAEP. *J. Educ. Stat.* 17:131–54

Miyazaki K, Hoshino T. 2009. A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika* 74(3):375–93

Monroe S, Cai L. 2014. Estimation of a Ramsay-curve item response theory model by the Metropolis–Hastings Robbins–Monro algorithm. *Educ. Psychol. Meas.* 74:343–69

Moustaki I, Knott M. 2014. Latent variable models that account for atypical responses *J. R. Stat. Soc. Ser. C* 63(2):343–60

Muraki E. 1992. A generalized partial credit model: application of an EM algorithm. *Appl. Psychol. Meas.* 16:159–76

Orlando M, Thissen D. 2000. New item fit indices for dichotomous item response theory models. *Appl. Psychol. Meas.* 24:50–64

Plieninger H, Meiser T. 2014. Validity of multi-process IRT models for separating content and response styles. *Educ. Psychol. Meas.* 74(5):875–99

Ramsay JO. 1991. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* 56(4):611–30

Rasch G. 1961. On general laws and the meaning of measurement in psychology. *Proc. 4th Berkeley Symp. Math. Stat. Probab.* 4:321–34

Raudenbush SW, Bryk AS. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage. 2nd ed.

Raudenbush SW, Liu X. 2000. Statistical power and optimal design for multisite randomized trials. *Psychol. Methods* 5(2):199–213

Reckase M. 2009. *Multidimensional Item Response Theory*. New York: Springer

Reise SP. 2012. The rediscovery of bifactor measurement models. *Multivar. Behav. Res.* 47(5):667–96

Reise SP, Waller NG. 2009. Item response theory and clinical measurement. *Annu. Rev. Clin. Psychol.* 5:27–48

Rijmen F. 2009. *Efficient full-information maximum likelihood estimation for multidimensional IRT models*. Tech. Rep. RR-09-03, Educ. Test. Serv., Princeton, NJ

Robbins H, Monro S. 1951. A stochastic approximation method. *Ann. Math. Stat.* 22:400–7

Rubin DB. 1981. The Bayesian bootstrap. *Ann. Stat.* 9:130–34

Rubin DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12:1151–72

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley

Rupp AA, Templin J, Henson RA. 2010. *Diagnostic Measurement: Theory, Methods, and Applications.* New York: Guilford

Samejima, F. 1969. Estimation of latent ability using a response pattern of graded scores. Psychom. Monogr. 17. Richmond, VA: Psychom. Soc.

Schilling S, Bock RD. 2005. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika* 70:533–55

Schmid J, Leiman JM. 1957. The development of hierarchical factor solutions. *Psychometrika* 22:53–61

Sinharay S, Johnson MS, Stern HS. 2006. Posterior predictive assessment of item response theory models. *Appl. Psychol. Meas.* 30(4):298–321

Sinharay S, Johnson MS, Williamson DM. 2003. *An application of a Bayesian hierarchical model for item family calibration*. ETS Rep. RR-03-04, Educ. Test. Serv., Princeton, NJ

Spybrook J, Raudenbush SW. 2009. An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educ. Eval. Policy Anal.* 31(3):298–318

Tatsuoka KK. 1983. Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20:345–54

Templin J, Henson RA. 2006. Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11:287–305

Thissen D, Cai L, Bock RD. 2010. The nominal categories item response model. In *Handbook of Polytomous Item Response Theory Models: Development and Applications*, ed. ML Nering, R Ostini, pp. 43–75. New York: Taylor & Francis

Thissen D, Steinberg L. 2009. Item response theory. In *Handbook of Quantitative Methods in Psychology*, ed. R Millsap, A Maydeu-Olivares, pp. 148–78. London: Sage

Thissen D, Wainer H, eds. 2001. *Test Scoring*. Hillsdale, NJ: Erlbaum

Thissen-Roe A, Thissen D. 2013. A two-decision model for responses to Likert-type items. *J. Educ. Behav. Stat.* 38:522–47

Thomas N, Gan N. 1997. Generating multiple imputations for matrix sampling data analyzed with item response models. *J. Educ. Behav. Stat.* 22(4):425–46

Thurstone LL. 1925. A method of scaling psychological and educational tests. *J. Educ. Psychol.* 16:433–51

van der Linden WJ. 2006. A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31:181–204

van der Linden WJ. 2008. Using response times for item selection in adaptive testing. *J. Educ. Behav. Stat.* 33:5–20

von Davier M. 2005. *A general diagnostic model applied to language testing data*. ETS Rep. RR-05-16, Educ. Test. Serv., Princeton, NJ

von Davier M, Gonzalez E, Mislevy R. 2009. What are plausible values and why are they useful? In *Issues and Methodologies in Large-Scale Assessments*, ed. M von Davier, D Hastedt, pp. 9–36. IERI Monogr. Ser. 2. Princeton, NJ: Inst. Econ. Res. Innov.

Wainer H, Bradlow ET, Wang X. 2007. *Testlet Response Theory and Its Applications*. Cambridge, UK: Cambridge Univ. Press

Wainer H, Vevea JL, Camacho F, Reeve BB, Rosa K, Nelson L. 2001. Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In *Test Scoring*, ed. D Thissen, H Wainer, pp. 343–87. Hillsdale, NJ: Erlbaum

Wirth RJ, Edwards MC. 2007. Item factor analysis: current approaches and future directions. *Psychol. Methods* 12:58–79

Woods CM, Thissen D. 2006. Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika* 71:281–301

Yang JS, Cai L. 2014. Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis–Hastings Robbins–Monro algorithm. *J. Educ. Behav. Stat.* 39:550–82

Annual Review of Statistics and Its Application

Volume 3, 2016

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics