Item Response Theory – A Statistical Framework for Educational and Psychological Measurement

Yunxiao Chen, Xiaoou Li, Jingchen Liu and Zhiliang Ying

Abstract. Item response theory (IRT) has become one of the most popular statistical models for psychometrics, a field of study concerned with the theory and techniques of psychological measurement. The IRT models are latent factor models tailored to the analysis, interpretation, and prediction of individuals' behaviors in answering a set of measurement items that typically involve categorical response data. Many important questions of measurement are directly or indirectly answered through the use of IRT models, including scoring individuals' test performances, validating a test scale, linking two tests, among others. This paper provides a review of item response theory, including its statistical framework and psychometric applications. We establish connections between item response theory and related topics in statistics, including empirical Bayes, nonparametric methods, matrix completion, regularized estimation, and sequential analysis. Possible future directions of IRT are discussed from the perspective of statistical learning.

Key words and phrases: Psychometrics, measurement theory, factor analysis, item response theory, latent trait, validity, reliability.

1. INTRODUCTION

Item response theory (IRT) models, also referred to as latent trait models, play an important role in educational testing and psychological measurement as well as several other areas of behavioral and cognitive measurement. Specifically, IRT models have been widely used in the construction, evaluation, and sometimes scoring, of large-scale high-stakes educational tests (e.g., Birdsall, 2011; Robin et al., 2014). Most national and international large-scale assessments for monitoring education quality, such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), which are of lower-stakes for test-takers, are also analyzed and reported under the IRT framework (Rutkowski et al., 2013). IRT models are a building block of student learning models for intelligent tutoring systems and personalized learning (Chen et al., 2018c; Khajah et al., 2014a,b; Tang et al., 2019; Wilson and Nichols, 2015). They also play an important role in the analysis

Yunxiao Chen is Assistant Professor, Department of Statistics, London School of Economics and Political Science, London, UK (Email: y.chen186@lse.ac.uk). Xiaoou Li is Assistant Professor, School of Statistics, University of Minnesota, Minneapolis, MN, USA. Jingchen Liu is Professor, Department of Statistics, Columbia University, New York, NY, USA. Zhiliang Ying is Professor, Department of Statistics, Columbia University, New York, NY, USA.

of health-related quality of life (Cella et al., 2002; Hays et al., 2000); specifically, these models have been the central statistical tool for the development of the Patient-Reported Outcomes Measurement Information System (PROMIS), a state-of-the-art assessment system for self-reported health that provides a standardized measurement of physical, mental, and social well-being from patients' perspectives (Cella et al., 2007). Furthermore, they are crucial to measuring psychological traits in various domains of psychology, including personality and psychopathology (Balsis et al., 2017; Reise and Waller, 2009; Wirth and Edwards, 2007). Besides these applications, IRT models receive wide applications in many other areas, such as political voting, marketing research, among others (e.g., Bafumi et al., 2005; De Jong et al., 2008).

IRT models are probabilistic models for individuals' responses to a set of items (e.g., questions), where the responses are typically categorical (e.g., binary/ordinal/nominal). These models are latent factor models from a statistical perspective, dating back to Spearman's factor model for intelligence (Spearman, 1904a). Some early developments on IRT include Richardson (1936), Ferguson (1942), Lawley (1943, 1944), among others. In the 1960s, Rasch (1960) and Lord and Novick (1968) laid the foundation of IRT as a theory for educational and psychological testing. Specifically, Rasch (1960) proposed what now known as the Rasch model, an IRT model with a very simple form that has important philosophical implications on psychological measurement and possesses good statistical properties brought by its natural exponential family form. Lord and Novick (1968) first introduced a general framework of IRT models and proposed several parametric forms of IRT models. In particular, the two-parameter (2PL) and three-parameter logistic (3PL) models (Birnbaum, 1968) were introduced that are still widely used in educational testing these days. Following these pioneer works, more flexible models and more powerful statistical tools have been developed to better measure human behaviors, promoting IRT to become one of the dominant paradigms for measurement in education, psychology, and related problems; see Carlson and von Davier (2017), Embretson and Reise (2000), and van der Linden (2018) for the history of IRT.

IRT models are closely related to linear factor models (see e.g., Anderson and Rubin, 1956; Bai and Li, 2012). The major difference is that linear factor models assume that the observed variables are continuous, while IRT models mainly focus on categorical variables. Due to their close connections, one can view IRT models as factor models for categorical data (Bartholomew, 1980). IRT models are also similar to generalized linear mixed models (GLMM; Berridge and Crouchley, 2011; Searle and McCulloch, 2001) in terms of the model assumptions, even though the two modeling frameworks are developed independently and focus on different inference problems. These connections will be further discussed in Section 2 below.

Classical test theory (CTT; Lord and Novick, 1968; Novick, 1966; Spearman, 1904b), also known as the true score theory, is another psychometric paradigm that was dominant before the prevalence of IRT. Perhaps the key advantage of IRT over CTT is that IRT takes item-level data as input and separately models and estimates the person and item parameters. This advantage of IRT allows for tailoring tests through judicious item selection to achieve precise measurement for individual test-takers (e.g., in computerized adaptive testing) or designing parallel

test forms with the same psychometric properties. It also provides mechanisms for placing different test forms on the same scale (linking and equating) and defining and evaluating test reliability and validity. These features of IRT will be discussed in the rest of the paper.

This paper reviews item response theory, focusing on its statistical framework and psychometric applications. In the review, we establish connections between IRT and related topics in statistics, such as empirical Bayes, nonparametric methods, matrix completion, regularized estimation, and sequential analysis. The purpose is three-fold: (1) to provide a summary of the statistical foundation of IRT, (2) to introduce some of the major problems in psychometrics to general statistical audiences through IRT as a lens, and (3) to suggest directions of methodological development for solving new measurement challenges in the big data era.

The rest of the paper is organized as follows. In Section 2, we provide a review of the statistical modeling framework of IRT and compare it with classical test theory and several related models. In Section 3, we discuss the statistical analyses under the IRT framework and their psychometric applications. In Section 4, we discuss several future directions of methodological development for solving new measurement challenges in the big data era. We end with a discussion in Section 5.

2. ITEM RESPONSE THEORY: A REVIEW AND RELATED STATISTICAL MODELS

2.1 Basic Model Assumptions

To be concrete, we discuss the modelling framework in the context of educational testing, though it is applicable to many other areas. We have N individuals responding to J test items. Let $Y_{ij} \in \{0, 1\}$ be a random variable representing test-taker i's response to item j, where $Y_{ij} = 1$ indicates a correct response and $Y_{ij} = 0$ otherwise. We further denote y_{ij} as a realization of Y_{ij} and denote $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iJ})^\top$ and $\mathbf{y}_i = (y_{i1}, ..., y_{iJ})^\top$. Note that we use boldface for vectors. An IRT model specifies the joint distribution of $\mathbf{Y}_i, i = 1, ..., N$.

An IRT model assumes that $\mathbf{Y}_i$, $i = 1, ..., N$, are independent, and models the joint distribution of random vector $\mathbf{Y}_i$ through the introduction of individual-specific latency. More precisely, a unidimensional IRT model assumes one latent variable for each individual i, denoted by θ_i, which is interpreted as the individual's level on a certain latent trait (i.e., ability) measured by the test. It is assumed that individuals' response patterns are completely characterized by their latent trait levels. This is reflected by the conditional distribution of $\mathbf{Y}_i$ given θ_i in an IRT model. The specification of this conditional distribution relies on two assumptions (1) a *local independence* assumption, saying that $Y_{i1}, ..., Y_{iJ}$ are conditionally independent given the latent trait level θ_i, and (2) an assumption on the item response function (IRF), also known as the item characteristic curve (ICC), defined as $g_j(\theta|\boldsymbol{\pi}_j) := P(Y_{ij} = 1|\theta_i = \theta)$, where $\boldsymbol{\pi}_j$ is a generic notation for parameters of item j. For example, $g_j(\theta|\boldsymbol{\pi}_j)$ takes the form

$$(2.1) \qquad \frac{\exp(\theta - b_j)}{1 + \exp(\theta - b_j)},$$

$$(2.2) \qquad \frac{\exp(d_j + a_j\theta)}{1 + \exp(d_j + a_j\theta)},$$

and

$$(2.3) \qquad c_j + (1 - c_j)\frac{\exp(d_j + a_j\theta)}{1 + \exp(d_j + a_j\theta)},$$

in the Rasch, 2PL, and 3PL models (Birnbaum, 1968; Rasch, 1960), respectively, where item parameters $\boldsymbol{\pi}_j = b_j$, (a_j, d_j), and (a_j, c_j, d_j), respectively. Probit (normal ogive) models, which were first proposed in Lawley (1943, 1944), are also commonly used (Embretson and Reise, 2000). These models replace the logit link in (2.1)-(2.3) by a probit link. That is, for example, the two-parameter probit model has the IRF $\Phi(d_j + a_j\theta)$, where Φ denotes the standard normal distribution. We also note that different items can have different IRFs. In summary, a unidimensional IRT model assumes that Y_{ij} is $g_j(\theta_i|\boldsymbol{\pi}_j)$ plus some noise, where the noise is often known as the measurement error. In educational testing, the IRF is usually assumed to be a monotonically increasing function (e.g., $a_j > 0$ in the 2PL and 3PL models) so that a higher latent trait level leads to a higher chance of correctly answering the item.

The complete specification of an IRT model remains to impose assumptions on the θ_i. According to Holland (1990b), θ_i can be viewed from two perspectives, known as the "stochastic subject" and the "random sampling" regimes, which lead to different parameter spaces. The "stochastic subject" regime was first adopted in the pioneering work of Rasch (1960). It treats each θ_i as an unknown model parameter to be estimated from data rather than a random sample from a certain population. This regime leads to the joint likelihood function (Lord, 1968)

$$(2.4) \qquad L_{JL}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, \theta_1, ..., \theta_N) = \prod_{i=1}^{N}\prod_{j=1}^{J} g_j(\theta_i|\boldsymbol{\pi}_j)^{y_{ij}}(1 - g_j(\theta_i|\boldsymbol{\pi}_j))^{1-y_{ij}}.$$

The "random sampling" regime assumes that θ_is are independent and identically distributed samples from a population with a density function f defined with respect to a certain dominating measure μ. As a result, one estimates the distribution function f from data, instead of the individual θ_is. This regime leads to the marginal likelihood function (Bock and Aitkin, 1981)

$$(2.5) \quad L_{ML}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, f) = \prod_{i=1}^{N}\left\{\int\prod_{j=1}^{J} g_j(\theta|\boldsymbol{\pi}_j)^{y_{ij}}(1 - g_j(\theta|\boldsymbol{\pi}_j))^{1-y_{ij}} f(\theta)\mu d\theta\right\}.$$

As discussed in Holland (1990b), both regimes have their unique strengths and weaknesses but the "random sampling" regime may be more solid as a foundation for statistical inference. The readers are referred to Holland (1990b) for a detailed discussion of the IRT assumptions from a statistical sampling viewpoint. We will revisit these likelihood functions in Section 3.1 when the corresponding estimation problems are discussed. Specifically, we will demonstrate that, under a double asymptotic regime where both the sample size N and the number of items J growing to infinity, the two likelihoods converge in a certain sense. Thus, the estimates obtained from one likelihood can approximate the other. As the "random sampling" regime is probably more widely accepted in the literature, we will adopt this regime in the rest of the paper, unless otherwise stated. We

also point out that, besides these two regimes, IRT models can be understood from a full Bayesian perspective or an "item sampling" rationale; see Thissen and Steinberg (2009) for further discussions.

Unidimensional IRT models discussed above can be viewed as a special case of multidimensional IRT models. In multidimensional IRT models, each individual is characterized by a vector of latent variables, denoted by $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{iK})^\top$, where K is the number of latent variables. The specification of a multidimensional IRT model is similar to that of a unidimensional model, except that g_j and f are multivariate. For example, the IRF of the multivariate two-parameter logistic (M2PL) model (Reckase, 2009) takes the form

$$(2.6) \qquad g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j) = \frac{\exp(d_j + a_{j1}\theta_1 + \cdots + a_{jK}\theta_K)}{1 + \exp(d_j + a_{j1}\theta_1 + \cdots + a_{jK}\theta_K)},$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^\top$ and $\boldsymbol{\pi}_j = \{a_{j1}, ..., a_{jK}, d_j\}$. In addition, the distribution f is often assumed to be multivariate normal (see e.g., Reckase, 2009). Recall that we denote vectors by boldface symbols. In the rest of the paper, we will use the boldface symbol $\boldsymbol{\theta}_i$ as a generic notation for the person parameters in a general IRT model, unless the discussion is specifically about unidimensional IRT models. Moreover, with slight abuse of notation, we denote the joint and marginal likelihoods for a general IRT model as $L_{JL}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$ and $L_{ML}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, f)$, respectively.

As commonly seen in latent variable models (e.g., linear factor models), suitable constraints are needed to ensure model identifiability. We take the M2PL model (2.6) as an example to illustrate this point, where $\boldsymbol{\theta}_i$ is assumed to follow a multivariate normal distribution. First, it is easy to observe that the model is not determined under location-scale transformations of the latent variables, in the sense that one can simultaneously transform the latent variables and the corresponding loading and intercept parameters without changing the distribution of observed data. This indeterminacy is typically resolved by letting each latent variable θ_{ik} have mean 0 and variance 1. Second, the model also has rotational indeterminacy, even after fixing the location and scale of the latent variables. That is, one can simultaneously rotate the latent variables $\boldsymbol{\theta}_i$ and the loading matrix $A = (a_{jk})_{J \times K}$ without changing the distribution of the response data. See Chapter 8, Reckase (2009) for a further discussion about these indeterminacies of IRT models.

The rotational indeterminacy is handled differently under the confirmatory and exploratory settings of IRT analysis. Under the confirmatory setting, design information (e.g., from the blueprint of a test) reveals the relationship between the items and the latent variables being measured. This design information can be coded by a $J \times K$ binary matrix, often known as the Q-matrix (Tatsuoka, 1983), $Q = (q_{jk})_{J \times K}$, where $q_{jk} = 1$ if item j directly measures the kth dimension and $q_{jk} = 0$ otherwise. The Q-matrix imposes zero constraints on the loading parameters. That is, a_{jk} will be constrained to 0 when $q_{jk} = 0$ so that Y_{ij} is conditionally independent of θ_{ik} given the rest of the latent traits. Figure 1 provides the path diagram of the multidimensional IRT model given its Q-matrix. As indicated by the directed edges from the latent traits to the responses, the two latent traits are directly measured by items 1-3 and 3-5, respectively. The Q-matrix thus takes
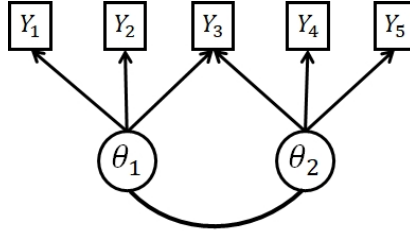
FIG 1. *the path diagram of multidimensional IRT model given Q-matrix in equation 2.7. The Q-matrix is indicated by the directed edges from the latent traits to the responses. The undirected edge means that the two latent traits are allowed to be correlated. The individual subscript i is omitted here for simplicity.*

the form

$$(2.7) \qquad Q = \left(\begin{array}{ccccc} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{array} \right)^{\top}.$$

For a carefully designed test whose Q-matrix satisfies suitable regularity conditions, it can be shown that the rotational indeterminacy no longer exists, as rotating the loading matrix will lead to violation of the zero constraints imposed by the Q-matrix (Anderson and Rubin, 1956; Chen et al., 2020a). Under the exploratory setting, the Q-matrix is not available. However, one would still assume, either explicitly or implicitly, that the relationship between the items and the latent traits can be characterized by a sparse Q-matrix and thus the corresponding loading matrix is sparse. Under this assumption, one fixes the rotation of the latent vector according to the sparsity level of the corresponding loading matrix. This problem plays an important role in the multidimensional measurement and will be further discussed in Section 3.4.

2.2 Summary of IRT Analysis

We describe statistical analyses under an IRT model and their purposes while leaving the more technical discussions to Section 3. We divide the analyses into the following four categories:

1. *Estimation of the item-specific parameters.* The estimate can reveal the psychometric properties of each item, and thus is often used to decide whether an item should be chosen into a measurement scale, monitor the change in items' psychometric properties over time (e.g., change due to the leakage of the item in an educational test that may be revealed by the change in the parameter estimate), among others. Estimation methods have been developed under the "random sampling" and "stochastic subject" regimes; see Section 3.1.
2. *Estimation of the person-specific parameters.* This problem is often referred to as the prediction of person parameters under the "random sampling" regime (Holland, 1990b) where they are viewed as random objects. This problem is closely related to the scoring of individuals in the measurement problem. For example, in a unidimensional IRT model for educational testing where all the IRFs are monotone increasing in θ, two estimates satisfying $\hat{\theta}_i > \hat{\theta}_{i'}$ implies that test-taker i is predicted to answer all items

more correctly than test taker i'. This estimate is often interpreted as an estimate of the test taker's ability. For another example, when a multidimensional IRT model is applied to a personality test, each component of the estimate $\hat{\boldsymbol{\theta}}_i$ may be interpreted as an estimate of the individual's level on a certain personality trait. As will be further discussed in Section 3.2, supposing that all the item parameters have been accurately estimated, the person parameters can be estimated based on the individual's responses to a subset of the items rather than all the test items. This is an important feature of IRT that is substantially different from the classical test theory; see a comparison between the two paradigms in Section 2.3 below.

3. *Model evaluation.* As the psychometric interpretations of an IRT model rely on the (approximate) satisfaction of the model assumptions, these assumptions need to be checked carefully, from the evaluation of individual assumptions (e.g., the forms of IRF and marginal distribution, local independence assumption, etc.) to assessing the overall goodness-of-fit. Many statistical methods have been developed for evaluating IRT models; see Section 3.3 for a discussion.

4. *Learning the latent structure of IRT models.* Like in exploratory factor analysis (e.g., Anderson, 2003), it is also often of interest to uncover the underlying structure of a relatively large number of items in multidimensional IRT models by learning the latent dimensionality and a sparse representation of the relationship between the latent variables and the items, i.e., the Q-matrix mentioned previously. The learning of the latent dimensionality is closely related to the determination of the number of factors in exploratory factor analysis. The Q-matrix learning problem is closely related to, and can be viewed as an extension of, the analytic rotation analysis (Browne, 2001) in exploratory factor analysis. Besides, the differential item functioning problem to be discussed in Section 2.8 also involves learning the relationship between the observed responses, latent traits, and individuals' covariate information. Further discussions can be found in Section 3.4.

2.3 Comparison with Classical Test Theory

We now compare IRT with classical testing theory (CTT; Lord and Novick, 1968), a more classical paradigm for measurement theory. The major difference is that CTT uses the test total score to measure individuals. It is suitable when different individuals answer the same items, but is less powerful, and sometimes infeasible, when individuals receive different test items (e.g., in computerized adaptive testing where different test takers receive different sets of items). More precisely, under the current notation, if all the items are equally weighted, then the total score of individual i is defined as $X_i := \sum_{j=1}^{J} Y_{ij}$. The CTT decomposes the total score as $X_i = T_i + e_i$, where T_i is the true score of person i and e_i is the measurement error of the current test administration satisfying $Ee_i = 0$. Conceptually, the true score T_i is defined as the expected number-correct score over an infinite number of independent administrations of the same test, hypothetically assuming that the individual does not keep the memory of test items after each administration (Chapter 2, Lord and Novick, 1968). CTT further assumes that T_i and e_i are uncorrelated and (T_i, e_i), $i = 1, ..., N$, are i.i.d. samples from a population. Under the CTT framework, the measurement of test-takers' ability

becomes to estimate T_is. A natural estimator of T_i is the total score X_i. This estimator is unbiased, since $\mathbb{E}(X_i|T_i) = T_i$.

A major contribution of CTT is to formally take the effect of measurement error into account in the modeling of testing data. This uncertainty leads to the concept of test reliability, defined as

$$\rho_{TX}^2 := \frac{Var(T_i)}{Var(X_i)} = 1 - \frac{Var(e_i)}{Var(X_i)},$$

which reflects the relative influence of the true and error scores on attained test scores. This coefficient is closely related to the coefficient of determination (i.e., R-squared) in linear regression. However, it is worth noting that based on a single test administration and without additional assumptions, one cannot disentangle the effects of the true score and the measurement error from the observed total scores. In other words, the true score T_i is not directly observed in CTT due to its latency, unlike the dependent variables in linear regression. Consequently, ρ_{TX}^2 cannot be estimated with the total scores only. In fact, methods for the estimation of test reliability, including Cronbach's alpha (Cronbach, 1951) and the split-half, test-retest, and parallel-form reliability coefficients (Chapter 9, Lord and Novick, 1968), all require additional assumptions. These additional assumptions essentially create repeated measurements of true score T_i. On the other hand, repeated measurements are automatically taken into account in an IRT model by modeling item-level data (each item as a repeated measure of the latent trait). Consequently, similar reliability coefficients are more straightforward to define and estimate under the IRT framework (Kim, 2012). For example, an analogous definition of ρ_{TX}^2 under a unidimensional IRT model is the so-called marginal reliability, defined as $Var(E(\theta_i|\mathbf{Y}_i))/Var(\theta_i)$.

The CTT and IRT frameworks are closely related to each other. In particular, CTT can be regarded as a first-order IRT model (Holland and Hoskens, 2003). That is, under a unidimensional IRT model given in Section 2.1, the true score of a test can be defined as

$$T_i := E(X_i|\theta_i) = \sum_{j=1}^{J} g_j(\theta_i|\boldsymbol{\pi}_j).$$

Under the monotonicity assumption of the IRFs, the true score T_i can be viewed as a monotone transformed latent trait level. The model parameters θ_i and $\boldsymbol{\pi}_j$ can be estimated from item-level data, leading to an estimate of the true score T_i.

2.4 Connection with Linear Factor Model

We next discuss the connection between IRT models and linear factor models. Specifically, we will show that these two types of models can be viewed as special cases of a general linear latent variable modeling framework (GLLVM, Chapter 2, Bartholomew et al., 2011), where IRT models focus on categorical items while linear factor models concern continuous variables. Moreover, it will be shown that by taking an underlying variable formulation (Chapter 4, Bartholomew et al., 2011), an IRT model can be viewed as a truncated version of the linear factor model.

2.4.1 A general linear latent variable modeling framework. The GLLVM framework was first introduced in Bartholomew (1984), and this framework has been further unified and extended in a wider sense in Moustaki and Knott (2000) and Rabe-Hesketh and Skrondal (2004). Recall that an IRT model consists of (1) specification of the conditional distribution of each response Y_{ij} given the latent trait, and (2) a local independence assumption and (3) an assumption on the marginal distribution of θ_i. The GLLVM specifies a general family of latent variable models following the three components (1)–(3). It allows for flexible choices of the conditional distribution of Y_{ij} given $\boldsymbol{\theta}_i$ in (1) and the marginal distribution of $\boldsymbol{\theta}_i$ in (3), where the latent variables $\boldsymbol{\theta}_i$ are allowed to be unidimensional or multidimensional. In this framework, the conditional distribution of Y_{ij} given $\boldsymbol{\theta}_i$ is allowed to be any generalized linear models (McCullagh and Nelder, 1989). Specifically, one obtains a linear factor model if Y_{ij} given $\boldsymbol{\theta}_i$ follows a normal distribution $N(d_j + a_{j1}\theta_1 + \cdots + a_{jK}\theta_K, \sigma_j^2)$ and $\boldsymbol{\theta}_i$ follows a multivariate normal distribution. In contrast, as reviewed previously, for binary response data, Y_{ij} given $\boldsymbol{\theta}_i$ follows a Bernoulli distribution with mean $g_j(\boldsymbol{\theta}_i|\boldsymbol{\pi}_j)$ in a (multidimensional) IRT model.

Most commonly used IRT models, including IRT models for other types of responses (e.g., ordinal/nominal) and multidimensional IRT models, can be viewed as special cases of the GLLVM. Such models include the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded response model (Samejima, 1969), the nominal response model (Bock, 1972), and their multidimensional extensions (Reckase, 2009). Under the GLLVM framework, an IRT model can be further embedded into a structural equation model, where the latent trait/traits measured by the IRT model can serve as latent dependent or explanatory variables in a structural equation model for studying the structural relationships among a set of latent and observed variables (Bollen, 1989).

2.4.2 Underlying variable formulation. Taking an underlying variable formulation (Christoffersson, 1975; Muthén, 1984), one can obtain a multidimensional IRT model by truncating a linear factor model. Suppose that $\tilde{\mathbf{Y}}_i = (\tilde{Y}_{i1}, ..., \tilde{Y}_{iJ})^\top$ follows a linear factor model with K factors. Let $Y_{ij} = 1_{\{\tilde{Y}_{ij} \geq 0\}}$. Then $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iJ})^\top$, which contains truncated $\tilde{Y}_{ij}$, follows a K-dimensional IRT model. More specifically, suppose that $\tilde{Y}_{ij}$ given $\boldsymbol{\theta}_i$ follows a normal distribution $N(d_j + a_{j1}\theta_1 + \cdots + a_{jK}\theta_K, 1)$. Then $P(Y_{ij} = 1|\boldsymbol{\theta}_i) = P(\tilde{Y}_{ij} \geq 0|\boldsymbol{\theta}_i) = \Phi(d_j + a_{j1}\theta_1 + \cdots + a_{jK}\theta_K)$. Recall that Φ is the cumulative distribution function of the standard normal distribution. Together with the assumptions of local independence and the marginal distribution f, it specifies a multidimensional IRT model under the probit link function. This formulation can be easily extended for ordinal response data.

Assuming normality in both the conditional distribution of $\tilde{\mathbf{Y}}_i$ given $\boldsymbol{\theta}_i$ and the marginal distribution of $\boldsymbol{\theta}_i$, the underlying variable formulation is closely related to the tetrachoric/polychoric correlations for measuring the association between binary/ordinal variables, which dates back to the seminal work of Pearson (1900a). As will be discussed in Section 3.1, this connection leads to a computationally efficient method for estimating the corresponding multidimensional IRT models.

2.5 Connection with Analysis of Contingency Tables

IRT models have a close relation with categorical data analysis, particularly the analysis of large sparse multidimensional contingency tables; see Fienberg (2000). In fact, response data for J items can be regarded as a J-way contingency table with 2^J cells, where each cell records the total count of a response pattern (i.e., a binary response vector). Note that the J-way contingency table is a sufficient statistic for the raw response data. This contingency table is typically sparse, as the sample size N is usually much smaller than the number of cells when the number of items J is moderately large (e.g., $J \geq 20$). Even when the sample size and the number of cells are comparable, some cells can still be sparse since the counts in the cells may be highly dependent on each other. Parsimonious models have been proposed for analyzing high-way contingency tables, which impose sparsity in the coefficients for higher-order interaction terms.

As pointed out in Holland (1990a), IRT models approximate the second-order loglinear model for J-way contingency tables

$$(2.8) \qquad P(\mathbf{Y}_i = \mathbf{y}) \propto \exp\left(\mathbf{d}^\top \mathbf{y} + \frac{1}{2}\mathbf{y}^\top A A^\top \mathbf{y}\right),$$

where $\mathbf{d} = (d_1, ..., d_J)^\top$ and $A = (a_{jk})_{J \times K}$ (K is often chosen to be much smaller than J, leading to a parsimonious model). This model does not explicitly contain latent variables. It is also known as an Ising model (Ising, 1925), an important model in statistical mechanics. Note that the same notations are used in the second-order loglinear model as those in the M2PL model in (2.6), for reasons explained below. That is, the second-order loglinear model can be viewed as an M2PL model with a special marginal distribution for the latent variables. More precisely, if the joint distribution of $(\mathbf{Y}_i, \boldsymbol{\theta}_i)$ takes the form

$$(2.9) \qquad f(\mathbf{y}, \boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} + \mathbf{y}^\top A \boldsymbol{\theta} + \mathbf{d}^\top \mathbf{y}\right),$$

then integrating out $\boldsymbol{\theta}$ gives the second-order log-linear model (2.8). Furthermore, it can be easily shown that the conditional distribution of $\mathbf{Y}_i$ given $\boldsymbol{\theta}_i$ is the same as that of the M2PL model given by equation (2.6). Under the joint model (2.9), the marginal distribution of $\boldsymbol{\theta}_i$ becomes a Gaussian mixture. For more discussions on the connection between IRT models and loglinear models, we refer readers to Fienberg and Meyer (1983), Ip (2002a), and Tjur (1982).

2.6 Comparison with Generalized Linear Mixed Models

The specification of IRT models under the "random sampling" regime is similar to that of generalized linear mixed models (GLMM; Berridge and Crouchley, 2011; Searle and McCulloch, 2001), and many IRT models can be viewed as special cases of GLMM. GLMM extends the generalized linear model to analyzing grouped (i.e., clustered) data commonly seen in longitudinal or repeated measures designs. A GLMM adds a random effect into a generalized linear model (e.g., logistic regression model) while keeping the fixed effect in the generalized linear model for studying the relationship between observed covariates and an outcome variable. The random effect is used to model the dependence among outcome variables within the same group (cluster). The IRT models introduced

in Section 2.1 can all be viewed as special cases under the GLMM framework. Specifically, each individual can be viewed as a group, and the individual's item responses can be viewed as repeated measures. The latent variables $\boldsymbol{\theta}_i$ correspond to the random effects in the GLMM. There are no observed covariates in these basic models but there are more complex IRT models to be discussed in Section 2.8 below that make use of individual-specific and item-specific covariates. Due to the similarities between IRT models and GLMMs, the estimation methods to be discussed in Section 3.1 for IRT models also apply to GLMMs.

While the two families of models largely overlap with each other in terms of the model assumptions, their main purposes are different, at least in the historical applications of these models. That is, the GLMM focuses on explaining data by testing certain hypotheses about the fixed effect parameters (i.e., the regression coefficients for observed covariates), treating the random effect as a component of the model that is not of interest in the statistical inference but necessary for capturing the within group-dependence. On the other hand, IRT models tend to focus on measuring individuals. Thus, the inference about the latent variables $\boldsymbol{\theta}_i$, the random effect from the GLMM perspective, is of particular interest. However, it is worth pointing out an important family of IRT models, called the explanatory IRT models (De Boeck and Wilson, 2004), that combines the explanatory perspective of GLMM and the measurement perspective of IRT. These models are specified under the GLMM framework, incorporating person- and item-specific covariates to explain the characteristics of individuals and items, respectively.

2.7 Connection with Collaborative Filtering and Matrix Completion

IRT models are also closely related to collaborative filtering (Koren and Bell, 2015; Zhu et al., 2016), a method of making automatic predictions (filtering) about the interests of a user by collecting preference/taste information from many users (collaborating) that is widely used in recommendation systems (e.g., e-commerce). The user-by-item matrix in collaborative filtering can be viewed as an item response data matrix, for which a large proportion of entries are missing due to the nature of the problem. A famous example of collaborative filtering is the Netflix challenge on movie recommendation (Feuerverger et al., 2012). Data of this challenge are ratings from a large number of users to a large set of movies, where many missing values exist as each user only watched a relatively small number of movies. The goal is to learn the preferences of each user on movies that they have not watched. The collaborative filtering problem, including the Netflix example, can be cast into a matrix completion problem that concerns filling the missing entries of a partially observed matrix. When the entries of the data matrix are binary or categorical-valued, the problem is known as a one-bit or categorical matrix completion problem.

Without further assumptions, the matrix completion problem is ill-posed since the missing entries can be assigned arbitrary values. To create a well-posed problem, a low-rank assumption is typically imposed to reduce the number of parameters, which is similar to the introduction of low-dimensional latent variables in IRT models. More precisely, for the completion of a binary or categorical matrix, it is often assumed that the data matrix follows a probabilistic model parameterized by a low-rank matrix (Bhaskar, 2016; Bhaskar and Javanmard, 2015; Cai and Zhou, 2013; Davenport et al., 2014; Zhu et al., 2016). From the sta-

tistical sampling perspective, these models take the "stochastic subject" regime that treats the user-specific parameters as fixed parameters. The specifications of such models are very similar to multidimensional IRT models. In particular, the one-bit matrix completion problem aims to recover the matrix $(p_{ij})_{N \times J}$, where $p_{ij} = P(Y_{ij} = 1|\boldsymbol{\theta}_i) = g_j(\boldsymbol{\theta}_i|\boldsymbol{\pi}_j)$. A major difference between collaborative filtering and psychometric applications of IRT is that psychometric applications are typically interested in the inference of the latent variables and the item parameters, while collaborative filtering only focuses on the inference of p_{ij}s. As will be discussed in Section 4.3, some psychometric problems may be viewed as collaborative filtering problems, and may be solved efficiently by matrix completion and related algorithms.

2.8 More Complex IRT Models and Their Psychometric Applications

In what follows, we review several more complex IRT models beyond the basic forms given in Section 2.1, and discuss their psychometric applications. Note that all these models can be viewed as special cases under the GLLVM framework discussed in Section 2.4.

2.8.1 IRT models involving covariates. Sometimes, covariates of the individuals are collected together with item response data. We use $\mathbf{x}_i$ to denote p-dimensional observed covariates of individual i. Covariate information can be incorporated into the IRT model in different ways for different purposes. We discuss two types of models often used in psychometrics.

First, the covariates may affect the distribution of the latent traits, but not directly on those of the item responses. Such models are known as the latent regression IRT models (Mislevy, 1984, 1985). They are useful in large-scale assessments, such as PISA and TIMSS, for estimating group-level distributions of the corresponding latent traits for policy-relevant sub-populations such as gender and ethnicity groups. For example, covariate information can be incorporated into a unidimensional IRT model (e.g., 2PL model) by assuming θ_i to follow a normal distribution $N(\mathbf{x}_i^\top \boldsymbol{\beta}, 1)$ instead of a standard normal distribution, where $\boldsymbol{\beta}$ is a p-dimensional vector of the coefficients. Note that the rest of the model assumptions remain the same (i.e., the IRFs and the local independence assumption). In particular, the covariates are not involved in the IRFs, so that the covariates do not directly affect the distribution of item responses. This model implies that the mean of the latent trait distribution depends on the covariates. When the covariates are indicators of group memberships (e.g., gender, ethnicity), the model allows the group means to be different. Figure 2 provides a graphical representation of the latent regression IRT model and other relevant models. Specifically, panels (a) and (b) of the figure show the path diagrams for a basic IRT model and a latent regression IRT model, respectively. In panel (b), the arrow from $\mathbf{x}$ to $\boldsymbol{\theta}$ implies that the distribution of $\boldsymbol{\theta}_i$ depends on the covariates and the absence of an arrow from $\mathbf{x}$ to the responses implies that the covariates do not directly affect the distribution of the response variables given the latent traits. The latent regression models can also be regarded as a special explanatory IRT model (De Boeck and Wilson, 2004); see Section 2.6 for a brief discussion of explanatory IRT models.

Second, the covariates may affect the distributions of the latent traits and the responses to some items. Such models are typically known as the Multiple Indi-
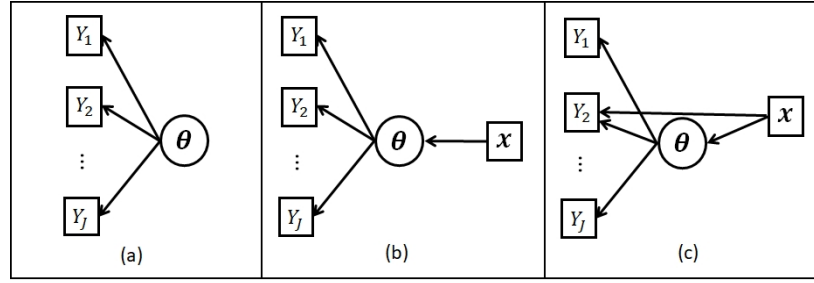
Fig 2. *Path diagrams for three IRT models, including (a) an IRT model without covariates, (b) a latent regression model, and (c) an MIMIC model. The individual subscript i is omitted here for simplicity.*

cators, Multiple Causes (MIMIC) models, originally developed in the literature of structural equation models (Goldberger, 1972; Robinson, 1974; Zellner, 1970) and then extended to IRT models (Muthén, 1985, 1988) for studying differential item functioning (DIF). DIF refers to the situation when items may function differently for different groups of individuals (e.g., gender/ethnicity) or they may measure even different things for members of one group as opposed to members of another. For example, a reading comprehension item in a language test may show DIF between the male and female groups if the content of the reading paragraphs is about a specific topic that either males or females have more experience with. The MIMIC model provides a nice way to describe the DIF phenomenon. For simplicity, consider the case of a single binary covariate $x_i \in \{0, 1\}$ that indicates the group membership. Then MIMIC model let the IRFs of the DIF items depend on the group membership, while allowing the two groups to have different latent trait distributions. For example, suppose that item j is the only DIF item among J items. Then a simple MIMIC model for DIF can be set as follows. The IRF of item j can be modeled as

$$(2.10) \qquad \frac{\exp(d_j + a_j\theta + \gamma_j x_i)}{1 + \exp(d_j + a_j\theta + \gamma_j x_i)},$$

which takes a 2PL model framework with γ_j being the parameter characterizing the group effect on the IRF. The rest of the items are DIF-free items, and thus, their IRFs still take the standard 2PL form (2.2). The latent trait distribution can be modelled by setting θ_i given x_i to follow $N(\beta x_i, 1)$. Panel (c) of Figure 2 gives the path diagram for a MIMIC model, where the second item is a DIF item. In real-world DIF analysis, DIF items are unknown and need to be detected based on data. As will be discussed in Section 3.4, the DIF analysis can be cast into a model selection problem. Comparing with other statistical model selection problems, such as variable selection in regression models, the current model has additional location-shift indeterminacy brought by the introduction of covariates, which complicates the analysis.

2.8.2 Discrete latent variables. The latent variables in an IRT model can also be discrete, though in many commonly used models, especially unidimensional models, they are assumed to be continuous variables. IRT models with categorical latent variables are known as latent class models (Goodman, 1974a,b; Haberman, 1977b; Lazarsfeld, 1950, 1959; Lazarsfeld and Henry, 1968). In the context of

psychological and educational measurement, a latent class model is preferred when the goal of measurement is to classify people into groups, while an IRT model with continuous latent variables is preferred if the goal is to assign people scores on a continuum. In particular, the diagnostic classification models (DCMs), also known as the cognitive diagnosis models, are a family of restricted latent class models dating back to the works of Haertel (1989) and Macready and Dayton (1977). DCMs have received much recent attention in educational assessment, psychiatric evaluation, and many other disciplines (Rupp et al., 2010; von Davier and Lee, 2019). In particular, such models have played an important role in intelligent tutoring systems (Cen et al., 2006; Doignon and Falmagne, 2012) to support personalized learning. In DCMs, an individual is represented by a latent vector $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{iK})^\top$, each component of which is a discrete latent variable that represents the individual's status on a certain attribute. For example, in the binary case, $\theta_{ik} = 1$ may represent the mastery of skill k, and $\theta_{ik} = 0$ otherwise. As a result, a diagnostic classification model classifies each individual given their item responses into a latent class that represents the individual's profile on multiple discrete attributes.

Measurement based on DCMs is multidimensional and confirmatory in nature, as these models are used to measure multiple fine-grained constructs of individuals simultaneously. As with the multidimensional IRT models discussed in Section 2.1, the Q-matrix also plays an important role in the DCMs. Recall that the Q-matrix takes the form $Q = (q_{jk})_{J \times K}$, where each entry q_{jk} is binary, indicating whether item j directly measures the kth dimension or not. Given the Q-matrix, different IRFs have been developed to capture the way a set of relevant attributes affect an item. For example, the Deterministic Inputs, Noisy "And" gate (DINA) model (Junker and Sijtsma, 2001) is a popular diagnostic classification model in educational measurement. The IRF of this model assumes that an individual will answer an item correctly provided that the individual acquires all the relevant skills, subject to noise due to slipping (when the individual can answer correctly) and guessing (when the individual is not able to answer correctly) behaviors. More precisely, the DINA model assumes binary latent variables (i.e., $\theta_k \in \{0, 1\}$) and the following IRF,

$$(2.11) \qquad g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j) = \begin{cases} 1 - s_j & \text{if } \theta_k \geq q_{jk}, \text{ for all } k = 1, ..., K, \\ g_j & \text{otherwise,} \end{cases}$$

where s_j and g_j are two item-specific parameters capturing the chances of answering incorrectly due to carelessness when he/she is able to solve the problem and of guessing the answer correctly, respectively, and $\boldsymbol{\pi}_j = (s_j, g_j)$. The indicating event in (2.11), $1_{\{\theta_k \geq q_{jk}, k=1,\cdots,K\}}$ is also known as the ideal response. Several DCMs have been developed to account for different types of psychological processes underlying item response behavior, such as the Deterministic Inputs, Noisy "Or" gate (DINO) model (Templin and Henson, 2006) and reparametrized unified models (DiBello et al., 1995; Junker and Sijtsma, 2001), for which different forms of IRFs are assumed. All these DCMs can be viewed as special cases of a general diagnostic classification model (de la Torre, 2011; Henson et al., 2009; von Davier, 2008), where the IRF takes the general form containing all the inter-

actions of latent variables

$$
\begin{aligned}
h_j^{-1}(g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j)) =& a_{j0} + a_{j1}q_{j1}\theta_1 + \cdots + a_{jK}q_{jK}\theta_K \\
& + a_{j12}q_{j1}q_{j2}\theta_1\theta_2 + \cdots + a_{j,K-1,K-2}q_{j,K-1}q_{j,K}\theta_{K-1}\theta_K \\
& + \cdots + a_{j,1,\cdots,K}\prod_{k=1}^{K}q_{jk}\theta_k.
\end{aligned}
$$

(2.12)

In (2.12), h_j is a link function (e.g., $h_j(x) = \exp(x)/(1 + \exp(x))$), and the item parameter vector $\boldsymbol{\pi}_j = \{a_{j0}, a_{j1}, \cdots, a_{j,1,\cdots,K}\}$. Note that the latent variable interactions are important to capture the disjunctive or conjunctive relationships between the latent attributes in the special cases such as the DINA and DINO models; see, e.g., de la Torre (2011) for more details.

Fundamental identifiability issues arise with the relatively more complex latent structure in DCMs. The interpretation and measurement of the latent constructs are only valid under an identifiable model. The identifiability problem of DCMs, and more generally of multidimensional IRT models, has two levels. First, under what measurement design are the model parameters ($\boldsymbol{\pi}_j$ and parameters in f, the marginal distribution of $\boldsymbol{\theta}$) identifiable, assuming that the Q-matrix is known and correctly specified? Second, if the Q-matrix is unknown, when can we simultaneously identify the model parameters and the Q-matrix from data? Efforts have been made to address these questions; see Chen et al. (2015), Chen et al. (2020a), Chiu et al. (2009), Fang et al. (2019), Fang et al. (2021), Gu and Xu (2020), Liu et al. (2013), Xu and Zhang (2016), Xu (2017), and Xu and Shang (2018).

2.8.3 Nonparametric IRT models. Nonparametric modeling techniques have been incorporated into IRT, yielding more flexible models. They play an important role in assessing the goodness-of-fit of parametric IRT models and providing robust measurement against model misspecification.

Under the "random sampling" regime, one can assume either the IRF $g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j)$ or the distribution f to be nonparametric. However, we note that model nonidentifiability generally occurs if assuming both to be nonparametric. In that case, one can simultaneously transform the IRFs and the latent variable distribution without changing the distribution of item responses; see Ramsay and Winsberg (1991) for a discussion. Even with only one of the IRFs and the marginal distribution f being nonparametric, model identifiability can still be an issue. This is because, for binary response data, the sufficient statistic (i.e., count of each of $2^J - 1$ response patterns) is of dimension $2^J - 1$, which does not grow for a fixed J. At the same time, there are infinitely many parameters in the nonparametric model component. See Douglas (2001) for a discussion about the identifiability of nonparametric IRT models and theory for the asymptotic identifiability when both the sample size N and the number of items J grow to infinity.

Cressie and Holland (1983) studied the identifiability of a semiparametric Rasch model, where the IRFs follow the form of a Rasch model and the marginal distribution f is nonparametric. For the same model, de Leeuw and Verhelst (1986) and Lindsay et al. (1991) further discussed the model identifiability and proposed a nonparametric marginal maximum likelihood estimator in the sense of Kiefer and Wolfowitz (1956), assuming f to be a mixture distribution with

unspecified weights at unknown points. Their results suggest that, under suitable conditions, the estimation of the item parameters is consistent under the asymptotic regime where J is fixed and N goes to infinity, even though there are more parameters than the sample size. Haberman (2005) extended the analysis by considering IRFs to take the 2PL and 3PL forms and found that the good properties of the nonparametric marginal maximum likelihood estimator for the semiparametric Rasch model do not carry over due to model non-identifiability. There have been other developments in the nonparametric modeling of item response functions. In these developments, the item response functions $g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j)$ are replaced by nonparametric functions. Generally speaking, research in this direction makes minimum assumptions on the item response functions, except for certain monotonicity or smoothness assumptions. Nonparametric function estimation methods have been applied to the estimation of nonparametric IRFs, such as spline methods; see Johnson (2007), Ramsay and Winsberg (1991), and Winsberg et al. (1984).

Many developments on non-parametric IRT have been made under the "stochastic subject" regime, either explicitly or implicitly, where the person parameters are treated as fixed parameters. Thus, only the IRFs are considered nonparametric. Specifically, Mokken scale analysis (Mokken, 1971; Mokkan and Lewis, 1982; Sijtsma and Molenaar, 2002), a pioneer work on non-parametric IRT, is developed under this regime. More precisely, a Mokken scale analysis model is a unidimensional IRT model, assuming that each IRF is a non-parametric monotonically nondecreasing function of the latent trait. Sometimes, it further makes the "non-intersecting IRFs" assumption that imposes a monotone ordering of IRFs. See Sijtsma and van der Ark (2017) for a discussion of these monotonicity assumptions. More general non-parametric IRT models have been proposed, for which theory and estimation methods have been developed. For an incomplete list of these developments, see Douglas (1997), Guo and Sinharay (2011), Johnson (2006), Ramsay and Abrahamowicz (1989), Sijtsma and Molenaar (2002), and Stout (1990). We point out that these "stochastic subject" models have similar identifiability issues as those under the "random sampling" regime. Therefore, restrictions on the model are needed to resolve the indeterminacies, and a double asymptotic regime is typically needed, i.e., both N and J growing to infinity, for establishing consistent estimation of the non-parametric IRFs (e.g. Douglas, 1997).

Finally, we point out that essentially all the existing works on non-parametric IRT models focus on unidimensional models. Non-parametric multidimensional IRT models remain to be developed, given that multidimensional measurement problems become increasingly more common these days. It is worth noting that deep autoencoder models, a family of deep neural network models, are essentially nonlinear and non-parametric factor models (Chapter 14, Goodfellow et al., 2016). These models are flexible and control the latent dimension through the architecture of the hidden layers in the deep neural network. In particular, the M2PL model given in Section 2.1 can be viewed as a special autoencoder model with one hidden layer. With this close connection, models and algorithms for deep autoencoders may be borrowed to develop non-parametric multidimensional measurement models.

3. STATISTICAL ANALYSIS UNDER IRT FRAMEWORK

In this section, we discuss the statistical analyses under the IRT framework as listed in Section 2.2.

3.1 Estimation of Item Parameters

We first consider the estimation of item parameters, which is often known as item calibration. This estimation problem is closely related to the estimation of fixed parameters in GLMMs and other nonlinear factor models, for which the methods reviewed below are also generally suitable. We discuss these estimation methods under the "random sampling" and "stochastic subject" regimes. As will be further discussed in Section 3.1.3 below, the two regimes converge in a certain sense when both the sample size and the number of items grow to infinity. In this double asymptotic sense, estimation under the "stochastic subject" regime can be viewed as an approximation to that under the "random sampling" regime, if one believes that the latter is more solid philologically as a foundation for statistical inference.

3.1.1 Marginal maximum likelihood estimation. The "random sampling" regime treats the person parameters as samples from a distribution. The marginal maximum likelihood (MML) estimator is the main estimation method under this regime and is also the most commonly used method in modern applications of item calibration. Suppose that the indeterminacies of an IRT model have been removed by imposing constraints on model parameters so that the model is identifiable. The MML simultaneously estimates the item parameters and the distribution of person parameters by maximizing the marginal likelihood function. That is,

$$(3.1) \qquad (\hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_J, \hat{f}) = \underset{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, f}{\arg\max} \log L_{ML}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, f).$$

Here, we assume the distribution f to take a parametric form. This estimator can be viewed as an empirical Bayes estimator (Efron and Morris, 1973; Robbins, 1956); see Efron (2003) and Zhang (2003) for a review of empirical Bayes methods.

It is worth emphasizing that linking is automatically performed in the MML estimator (3.1), as well as some of the estimators reviewed later including the JML estimator. Suppose that each individual i is only given a subset of the test items. This design is commonly used in large-scale assessments such as PISA and TIMSS to achieve good content coverage without requiring each individual to answer too many items. That is, let $B_i \subset \{1, ..., J\}$ be the set of items assigned to individual i. Then the marginal likelihood function can be written as

(3.2)

$$L_{ML}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, f) = \prod_{i=1}^N \left\{ \int \prod_{j \in B_i} g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j)^{y_{ij}} (1 - g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j))^{1-y_{ij}} f(\boldsymbol{\theta}) \mu d\boldsymbol{\theta} \right\},$$

When there is sufficient overlap between the subsets B_i and given the standard identifiability constraints for the IRT model, the item parameters and marginal distribution f can still be identified based on the marginal likelihood (3.2). Consequently, the item parameter estimates from the MML estimator based on (3.2) automatically lie on the same scale. We refer the readers to Mislevy and Wu

([1996](#)) for further discussions on the application of IRT models to item response data with missing values.

The optimization problem ([3.1](#)) is often solved by an Expectation-Maximization (EM) algorithm ([Bock and Aitkin](#), [1981](#); [Dempster et al.](#), [1977](#)). When a multidimensional IRT model is fitted, the EM algorithm for solving ([3.1](#)) is usually slow. This is because the algorithm has to frequently solve K-dimensional numerical integrals, whose complexity increases exponentially with K. To speed up the EM algorithm, stochastic versions of the EM algorithm have been proposed. These algorithms avoid the numerical integration by Monte Carlo simulation ([Cai](#), [2010a,b](#); [Diebolt and Ip](#), [1996](#); [Ip](#), [2002b](#); [Meng and Schilling](#), [1996](#); [Zhang et al.](#), [2020](#); [Zhang and Chen](#), [2020](#)). Among these developments, we draw attention to the stochastic approximation methods, also known as the stochastic gradient descent methods, proposed in [Cai](#) ([2010a,b](#)) and [Zhang and Chen](#) ([2020](#)), which date back to the seminal work of [Robbins and Monro](#) ([1951](#)) on stochastic approximation and the work of [Gu and Kong](#) ([1998](#)) that combines Markov chain Monte Carlo (MCMC) sampling and stochastic approximation for estimating latent variable models.

We now demonstrate how the optimization problem ([3.1](#)) can be solved by stochastic approximation. To simplify the notation, we use Ξ to denote all the fixed parameters in $\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J$ and f, and write the log marginal likelihood function as $l(\Xi)$. We further denote

$$l_i(\boldsymbol{\theta}_i, \Xi) = \log\left(\prod_{j=1}^{J} g_j(\boldsymbol{\theta}_i|\boldsymbol{\pi}_j)^{y_{ij}}(1 - g_j(\boldsymbol{\theta}_i|\boldsymbol{\pi}_j))^{1-y_{ij}} f(\boldsymbol{\theta}_i)\right)$$

as the log complete-data likelihood for individual i that is based on the joint distribution of $\boldsymbol{\theta}_i$ and $\mathbf{Y}_i$. Then it can be shown that the gradient of $l(\Xi)$ takes the form

$$\nabla l(\Xi) = \sum_{i=1}^{N} E\left(\frac{\partial l_i(\boldsymbol{\theta}_i, \Xi)}{\partial \Xi}\Big|\mathbf{Y}_i = \mathbf{y}_i\right),$$

where the conditional expectation is with respect to $\boldsymbol{\theta}_i$ given $\mathbf{Y}_i$. With this observation, the stochastic approximation methods ([Cai](#), [2010a,b](#); [Zhang and Chen](#), [2020](#)) for solving ([3.1](#)) iterate between two steps: (1) sample $\boldsymbol{\theta}_i$ from its conditional distribution given $\mathbf{Y}_i$, where the conditional distribution is based on $\Xi^{(t)}$, the current value of Ξ, and (2) given the obtained samples $\boldsymbol{\theta}_i^{(t)}$, $i = 1, ..., N$, update the value of Ξ by a stochastic gradient ascent, where the stochastic gradient at $\Xi^{(t)}$ is given by

$$\sum_{i=1}^{N} \frac{\partial l_i(\boldsymbol{\theta}_i^{(t)}, \Xi)}{\partial \Xi}\Big|_{\Xi=\Xi^{(t)}}$$

whose conditional expectation (given observed responses) is $\nabla l(\Xi)$. For multidimensional IRT models with many latent traits, it may not be straightforward to sample $\boldsymbol{\theta}_i$ from the conditional distribution, and MCMC methods are needed to perform the sampling step. Under mild conditions, $\Xi^{(t)}$ is guaranteed to converge to the solution of optimization problem ([3.1](#)), even when the samples in the sampling step are approximated by an MCMC algorithm ([Zhang and Chen](#), [2020](#)). Typical of stochastic approximation, the performance of these algorithms is sensitive to the step size in the stochastic gradient ascent step, where the step size

is required to decay to 0 at a suitable rate to ensure convergence. Cai (2010a,b) suggested to set the step size to delay at the rate $1/t$, which is known to be asymptotically optimal for the Robbins-Monro algorithm (Chung, 1954; Lai and Robbins, 1979). However, the $1/t$ rate is well-known to yield unstable results in practice as it decays to zero too fast. Zhang and Chen (2020) suggested to use a slower-decaying step size and the Polyak–Ruppert averaging procedure (Polyak and Juditsky, 1992; Ruppert, 1988) to improve the empirical performance of the stochastic approximation algorithm while maintaining a fast theoretical convergence rate.

3.1.2 Limited-information estimation methods. Alternative estimation methods are developed under the "random sampling" regime to bypass the high-dimensional integrals in the MML. These methods are known as limited-information methods. They do not consider the complete joint contingency table of all items, but only marginal tables up to a lower order (e.g., two-way tables based on item pairs).

We divide these methods into two categories. The first concerns the probit models discussed in Section 2.4.2 where the IRFs take a probit form, and marginal distribution of the latent traits is also a multivariate normal distribution. Making use of the underlying variable formulation, it can be shown that these IRT models can be estimated by first estimating the multivariate normal distribution of the underlying variables and then recover the IRT parameters based on the estimated distribution of the underlying variables. Note that the first step can be done efficiently using the one-way and two-way tables based on all the individual items and item pairs (Muthén, 1984), and the second step solves an optimization problem with no integrals involved. Consequently, this method can computationally efficiently estimate IRT models with many latent traits, especially when the number of items is not too large. Developments in this direction include Christoffersson (1975), Lee et al. (1990, 1992, 1995), and Muthén (1978, 1984).

The second category of methods makes use of the composite likelihood method (Besag, 1974; Cox and Reid, 2004; Lindsay, 1988). These methods allow for more general forms of the IRFs but still require the marginal distribution of the latent traits to be normal. More specifically, the fixed parameters are estimated by maximizing a composite likelihood based on the lower-order marginal tables. In particular, the pairwise likelihood is most commonly used that is constructed based on item pairs, though more general composite likelihood functions can be constructed based on triplets or quadruplets of items; see Jöreskog and Moustaki (2001), Katsikatsou et al. (2012), Vasdekis et al. (2012), and Vasdekis et al. (2014).

Although these limited-information methods are computationally less demanding than the MML approach, there are some limitations. First, as mentioned previously, these methods only apply to some restricted classes of IRT models (e.g., probit IRFs, multivariate normal distribution for the latent traits). They thus are not as generally appliable as the MML approach. Second, when the IRT model is correctly specified, the limited-information methods suffer from some information loss and thus are statistically less efficient than the MML estimator.

3.1.3 Joint maximum likelihood estimation. The joint maximum likelihood (JML) estimator refers to the estimation method that simultaneously estimates the item and person parameters by maximizing the joint likelihood function introduced in Section 2.1. This approach was first suggested in Birnbaum (1968), and has been used in item response analysis for many years (Lord, 1980; Mislevy and Stocking, 1989; Wood et al., 1978) until the MML approach becomes dominant. Since the joint likelihood function does not involve integrals, the computation of the JML estimator is typically much faster than that of the corresponding MML estimator, especially when the latent dimension is high. Despite the computational advantage, the JML estimator is still less preferred to the MML for item calibration, possibly due to two reasons. The first one is philosophical. As pointed earlier, the JML estimator naturally fits the "stochastic subject" regime, which, however, is less well accepted than the "random sampling" regime. The second reason is that the JML estimator lacks desirable asymptotic properties. Under the standard asymptotic regime where the number of item is fixed and the number of people goes to infinity, the JML estimation of the item parameters is inconsistent. This inconsistency is due to that the sample size and the dimension of parameter space grow at the same speed. This phenomenon was originally noted in Neyman and Scott (1948) under a linear model and further discussed by Andersen (1970) and Ghosh (1995) under IRT models.

While these reasons are valid under a setting when J is small, they may no longer be a concern under a large-scale setting when both the sample size N and the number of items J are large. We provide some justifications for joint-likelihood-based estimation, using the double asymptotic regime that both N and J grow to infinity. We first point out that the two likelihood functions tend to approximate each other under this regime. To see this, we do local expansion of the marginal likelihood function at the MML estimator $(\hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_J, \hat{f})$. Under the double asymptotic regime and by making use of the Laplace approximation for integrals, $\log L_{ML}(\hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_J, \hat{f})$ can be approximated by $\log L_{JL}(\hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_J, \hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_N)$ plus some smaller-order terms, where $\hat{\boldsymbol{\pi}}_j$ is the same MML estimator and $\hat{\boldsymbol{\theta}}_i$ is given by

$$\hat{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}} \ \log \left(\prod_{j=1}^{J} g_j(\boldsymbol{\theta}|\hat{\boldsymbol{\pi}}_j)^{y_{ij}} (1 - g_j(\boldsymbol{\theta}|\hat{\boldsymbol{\pi}}_j))^{1-y_{ij}} \right).$$

See Huber et al. (2004) for more details about this expansion.

We further point out that some notion of consistency can be established for the JML approach under the double asymptotic regime. Specifically, Haberman (1977a) showed that person and item parameters of a Rasch model can be consistently estimated when N and J grow to infinity at a suitable rate. Chen et al. (2021) extended the analysis of Haberman (1977a) under a setting that many entries of the response data matrix are missing. Chen et al. (2019, 2020a) considered a suitably constrained JML estimator for more general multidimensional IRT models and showed that the estimator achieves the optimal rate under this asymptotic regime. Specifically, the constrained JML estimator solves a JML

problem with constraints on the magnitudes of person and item parameters

$$(3.3) \quad (\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_N, \hat{\boldsymbol{\pi}}_1, ..., \hat{\boldsymbol{\pi}}_J) = \underset{\boldsymbol{\theta}_1, \cdots \boldsymbol{\theta}_N, \boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J}{\arg \max} \, L_{JL}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$$

$$s.t. \, \|\boldsymbol{\theta}_i\| \leq C, \|\boldsymbol{\pi}_j\| \leq C, i = 1, ..., N, j = 1, ..., J,$$

where C is a pre-specified constant. Under mild conditions and assuming a fixed latent dimension K, Chen et al. (2020a) showed that

$$(3.4) \quad \frac{\|\hat{P} - P^*\|_F}{\sqrt{NJ}} = O_p\left(\frac{1}{\sqrt{N \wedge J}}\right)$$

is the optimal rate (in minimax sense) for estimating P^*. Here, $\hat{P} = (g_j(\hat{\boldsymbol{\theta}}_i|\hat{\boldsymbol{\pi}}_j))_{N \times J}$ and $P^* = (g_j(\boldsymbol{\theta}_i^*|\boldsymbol{\pi}_j^*))_{N \times J}$, where $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\pi}}_j$ are from (3.3), and $\boldsymbol{\theta}_i^*$ and $\boldsymbol{\pi}_j^*$ are the true parameter values that are required to satisfy the constraints in (3.3). By making use of (3.4) and by proving extensions of the Davis-Kahan-Wedin sine theorem (Davis and Kahan, 1970; Wedin, 1972) for the perturbation of eigenvectors, it can be shown under suitable conditions that

$$(3.5) \quad \sqrt{\frac{\sum_{j=1}^J \|\hat{\boldsymbol{\pi}}_j - \boldsymbol{\pi}_j^*\|^2}{J}} = O_p\left(\frac{1}{\sqrt{N \wedge J}}\right).$$

Under the typical setting where N is much larger than J, the rates in (3.4) and (3.5) both become $1/\sqrt{J}$; that is, the accuracy of the JML estimator is mainly determined by the number of items. We conjecture that $\max_{j=1,...,J} \|\hat{\boldsymbol{\pi}}_j - \boldsymbol{\pi}_j^*\| = o_p(1)$, as both N and J grow to infinity. This conjecture may be proved by a careful expansion of the joint likelihood function at the constrained JML estimator.

With the above justifications, one may use the JML estimator as an approximation to the MML estimator in large-scale applications, when both N and J are large and the computation of the MML estimator is intensive.

3.1.4 Full Bayesian estimation. Full Bayesian methods have also been developed for the estimation of IRT models. These methods regard the unknown parameters in $\boldsymbol{\pi}_j$ and f as random variables and impose prior distributions for them. The estimation and associated uncertainty quantification are obtained based on the posterior distributions of the unknown parameters. MCMC algorithms have been developed for the full Bayesian estimation of different IRT models, especially multidimensional IRT models; see, for example, Culpepper (2015), Edwards (2010) and Jiang and Templin (2019).

3.1.5 Other estimation methods. We summarize two other estimation methods commonly used in IRT analysis. One is the conditional maximum likelihood, which essentially follows the "stochastic subject" regime. This method makes use of a conditional likelihood function that does not contain any person parameters. This estimator is not flexible enough in that the construction of the conditional likelihood relies heavily on the raw total score being a sufficient statistic for the person parameter and thus is only applicable to models within the Rasch family (Andersen, 1970, 1977; Andrich, 2010; Rasch, 1960).

A two-step procedure is typically used to analyze unidimenisional non-parametric IRT models that assume monotone IRFs (Douglas, 1997; Guo and Sinharay, 2011;

Johnson, 2006). In the first step, one estimates the individuals' latent trait level based on their total scores. Note that this step only makes sense for unidimensional IRT models with monotone IRFs. Then in the second step, the IRFs are estimated using a non-parametric regression procedure that regresses the response to each item on the estimated latent trait level. To obtain consistency in estimating the IRFs, one typically needs the measurement error in the first step to decay to 0, which requires the number of items J to grow to infinity. See Douglas (1997) for the asymptotic theory.

3.2 Estimation of Person Parameters

Here we discuss the estimation of person parameters, assuming that item parameters are known. This problem is closely related to the scoring of individuals in psychological and educational measurement. Two settings will be discussed, an offline setting for which response data have already been collected and an online setting for which responses are collected sequentially in real-time.

3.2.1 Offline estimation. We first consider the offline setting for the estimation of person parameters, assuming that the item parameters are known. This problem is similar to that of a regression problem. Specifically, the likelihood function of person parameters factorizes into a product of the likelihood functions of individual $\boldsymbol{\theta}_i$s. As a result, $\boldsymbol{\theta}_i$ can be estimated by the maximum likelihood estimator, where the likelihood function of $\boldsymbol{\theta}_i$ is given by

$$(3.6) \qquad L_i(\boldsymbol{\theta}_i) = \prod_{j=1}^{J} g_j(\boldsymbol{\theta}_i|\boldsymbol{\pi}_j)^{y_{ij}}(1 - g_j(\boldsymbol{\theta}_i|\boldsymbol{\pi}_j))^{1-y_{ij}}.$$

When the distribution f is specified, $\boldsymbol{\theta}_i$ can also be estimated by a Bayesian method, such as the maximum a posteriori probability (MAP) or expected a posteriori (EAP) estimator. In fact, the posterior distribution of θ_i is proportional to $L_i(\boldsymbol{\theta}_i)f(\boldsymbol{\theta}_i)$.

The estimation of $\boldsymbol{\theta}_i$ leads to the scoring of individuals. Specifically, in a unidimensional IRT model for an educational test, the estimates $\hat{\theta}_i$ provide a natural ranking of the test-takers in terms of their proficiency on the construct measured by the test. Note that $\boldsymbol{\pi}_j$ and f may be estimated using the procedures described in Section 3.1 above and then treated as known for the estimation of the person parameters. We also point out that some methods described in Section 3.1 automatically provide person parameter estimates as byproducts, such as the JML estimator, and the Bayesian estimator based on an MCMC algorithm in which posterior samples of the person parameters are obtained.

One advantage of scoring by IRT is that linking is automatically performed through the calibration of the items so that the item parameters are on the same scale as that of the latent traits. As a result, the estimated person parameters are aligned on the same scale, even when different students receiving different subsets of the items. In that case, for each person i, a likelihood function similar to that of (3.6) can be derived, where items involved in the likelihood function are the ones that person i receives. This property of IRT leads to an important application of IRT, the computerized adaptive testing (CAT). In CAT, a pool of items (also known as an item bank) is pre-calibrated, and test-takers receive different sets of items following a sequential item selection design. CAT applies an

online (i.e., sequential) method for estimating person parameters, which is made possible by techniques from sequential analysis, a branch of statistics concerned with adaptive experimental design and optimal stopping. We discuss this problem below.

3.2.2 Online estimation via computerized adaptive testing. IRT also serves as the foundation for computerized adaptive testing (CAT), a computer-based testing mode under which items are fed to an individual sequentially, adapting to the current knowledge about the individual's latent traits. In educational testing, the use of CAT avoids giving capable test-takers too many easy items and giving less capable test-takers too many difficult ones. Consequently, CAT can lead to accurate measurement of the individuals with a smaller number of items, in comparison with non-adaptive testing. The concept of adaptive testing was originally conceptualized by Lord (1971) in his attempt to apply the stochastic approximation algorithm of Robbins and Monro (1951) to design more efficient tests. This idea of adaptive testing was realized in early works on CAT, including Lord (1977) and Weiss (1974, 1978, 1982). A comprehensive review of the practice and statistical theory of CAT can be found in Chang (2015), van der Linden and Glas (2010) and Wainer et al. (2000).

The CAT problem can be regarded as a sequential experimentation and estimation problem, where an IRT model is assumed with known item parameters and continuous latent traits. The aim of CAT is to achieve a pre-specified level of accuracy in estimating each person parameter with a test length as short as possible. A CAT algorithm has three building blocks, (1) a stopping rule which decides whether to stop testing or not at each step based on the individual's current performance (i.e., performance on the finished items), (2) an item selection rule, which decides which item to give to the test-taker in the next step, and (3) an estimator of the individual's latent traits $\boldsymbol{\theta}$ and its inference.

There are statistical problems arising from the above CAT procedure. First, given a stopping rule and an item selection rule, how do we obtain $\hat{\boldsymbol{\theta}}$ and its standard error? The setting is different from estimating person parameters based on static item response data, as the responses are no longer conditionally independent given the latent trait level due to sequential item selection. This problem was investigated in Chang and Ying (2009), where $\hat{\boldsymbol{\theta}}$ can be obtained by solving a score equation. Theoretical results of consistency, asymptotic normality, and asymptotic efficiency were established by making use of martingale theory.

Second, how should the item selection rule and stopping rule be designed? This problem is closely related to the early stopping problem and the sequential experimental design problem in the literature of sequential analysis, where the former dates back to Wald's pioneer works on sequential testing (Wald, 1945; Wald and Wolfowitz, 1948) and the latter dates back to Chernoff's seminal works on sequential experimental design (Chernoff, 1959, 1972). These problems are major topics of sequential analysis; see Lai (2001) for a comprehensive review. More specifically, the optimal sequential decision on early stopping and item selection can be formulated under the Markov decision process (MDP) framework (Puterman, 2014), a unified probabilistic framework for optimal sequential decisions. In this MDP, the goal is to minimize a certain loss function (or equivalently, to maximize a certain utility function) that concerns both the accuracy of measurement and the number of items being used, with respect to early stopping and

item selection as possible actions that need to be taken at each step based on the currently available information.

A seemingly standard MDP problem, the item selection and early stopping problems in CAT typically cannot simply be solved by dynamic programming, the standard method for MDPs, due to the huge state space and action space. In fact, obtaining an exact optimal solution under a nonasymptotic setting is NP-hard under the CAT setting. Therefore, the developments of CAT procedures are usually guided by asymptotic analysis, heuristics of finite sample performance, and also practical constraints such as speed of the algorithm, item exposure, and balance of contents. For a practical solution to the CAT problem under unidimensional IRT models, Lord (1980) proposed to select the next item as the one with the maximum Fisher information at the current point estimate of θ, and Chang and Ying (1996) proposed methods based on a Kullback–Leibler (KL) information index, which takes uncertainty in the point estimate of θ into account under a Bayesian setting. Although these maximum information item selection methods are asymptotically efficient, they often do not perform well at the early stage of a CAT when the estimation of θ is inaccurate. This is because, an item with maximum information at the estimated θ may not be informative at the true value of θ when the true value and its estimate are far apart at the early stage. In addition, the use of maximum information item selection methods could lead to skewed item exposure rates. That is, some items could be frequently used in a CAT whereas others might never be used, which may lead to item leakage. To improve item selection at the early stage of a CAT, Chang and Ying (1999) proposed a multistage item selection method. This method stratifies the item pool into less discriminating items and discriminating items, where a discriminating item tends to have a large information index (e.g., Fisher information) at a certain θ value while a less discriminating item has a relatively smaller information index for all values of θ. It uses less discriminating items early in the test when estimation is inaccurate, and saves highly discriminating items until later stages. The stopping of a CAT procedure is often determined by the asymptotic variance of the θ estimate (Weiss and Kingsbury, 1984) and the sequential confidence interval estimation (Chang, 2005).

CAT methods have also been developed under DCMs where the latent variables are discrete. The CAT problem becomes a sequential classification problem under this setting, which is different from sequential estimation. New methods have been developed for the item selection, early stopping, and making final classification. See Cheng (2009), Liu et al. (2015), Tatsuoka and Ferguson (2003), and Xu et al. (2003).

The past decade has seen the increasing use of computerized multistage testing (Yan et al., 2016) in the educational testing industry, a testing mode that can be viewed as a special case of the CAT. Instead of adaptively selecting individual items, computerized multistage testing divides a test into multiple stages and adaptively selects a group of items for each stage based on an individual's previous performance. Instead of solving a standard sequential design problem as in CAT, computerized multistage testing involves solving a group sequential design problem. We note that similar group sequential design problems have been widely encountered in clinical trials, for which statistical methods and theory have been developed; see, for example, Chapter 4, Bartroff et al. (2012).

3.3 Evaluation of IRT Models

The psychometric validity of an IRT model relies on the extent to which its assumptions hold. In what follows, we discuss the evaluation of the overall goodness-of-fit of an IRT model and the assessment of specific model assumptions.

3.3.1 Overall goodness-of-fit. Assessing the overall goodness-of-fit of an IRT model can be cast into testing the null hypothesis of data being generated by the IRT model. In principle, this problem can be solved by Pearson's chi-squared test (Pearson, 1900b) given data with a large sample size. However, as mentioned in Section 2.5, the J-way contingency table is typically sparse when the number of items J is moderately large, resulting in the failure of the asymptotic theory for Pearson's chi-squared test.

There are two types of methods that give valid statistical inference for sparse contingency tables. The first type is bootstrap methods (e.g., Collins et al., 1993; Bartholomew and Tzamourani, 1999), which is typically time consuming. The second method is based on assessing the fit of lower-way marginal tables, such as two- or three-way tables based on item pairs or triplets, rather than a complete table with 2^J cells. Asymptotic theory holds for these marginal tables since they have much smaller numbers of cells. Developments in this direction include Bartholomew and Leung (2002), Christoffersson (1975), Maydeu-Olivares and Joe (2005, 2006) and Reiser (1996).

When an overall goodness-of-fit test suggests a lack of fit, it is desirable to obtain information about which specific assumptions are being violated. Methods for assessing individual assumptions of an IRT model will be discussed below. In addition, as statistical models are only an approximation to the real data generation mechanism, IRT models are typically found to lack fit when applying to real-world item response datasets that have a reasonably large sample size (Maydeu-Olivares and Joe, 2006).

3.3.2 Dimensionality. IRT models are often used in a confirmatory manner where the number of latent traits is pre-specified, especially for unidimensional IRT models. A question that needs to be answered is whether the pre-specified latent dimension is sufficient or some extra dimensions are needed. It is natural to answer this question by the comparison of IRT models with different numbers of latent traits. For example, to assess unidimensional assumption of the 2PL model, we may compare it with an M2PL model that has two or more latent traits. While seemly a simple problem of comparing nested models, the asymptotic reference distribution for the corresponding likelihood ratio test is not a chi-squared distribution due to the null model lying at boundary points or singularities of the parameter space in this problem (Chen et al., 2020b). This asymptotic reference distribution can be derived using a more general theory for the likelihood ratio test statistic (Chernoff, 1954; Drton, 2009).

Alternatively, one may also use an exploratory approach to directly learn the latent dimension from data, and then compare it with the pre-specified dimension. See Section 3.4 for further discussions.

3.3.3 Local independence. The local independence assumption plays an essential role in IRT models. It is closely related to the dimensionality assumption that is discussed in Section 3.3.2, in the sense that the existence of extra latent

dimensions can cause the violation of the local independence assumption. The local independence assumption is often assessed through an analysis of the residual dependence given the hypothesized latent traits. The goal is to find dependence patterns in data that are not attributable to the primary latent dimensions. Such patterns could reveal how the hypothesized latent structure is violated, and to what extent the violation is.

One type of methods is based on the residuals for lower-way marginal tables (e.g., Chen and Thissen, 1997; Liu and Maydeu-Olivares, 2013; Yen, 1984). These methods search for subsets of items (e.g., item pairs) that violate the local independence assumption, by defining residual-based diagnostic statistics and testing the goodness-of-fit of the corresponding lower-way marginal tables based on these statistics. Although these methods have reasonably good power according to simulation studies, there is a gap. That is, the null hypothesis in the hypothesis tests based on different marginal tables is always "the fitted IRT model holds for all items", rather than "local independence holds within the corresponding subset of items". Therefore, from the perspective of hypothesis testing, when the null hypothesis is rejected for a marginal table, it does not directly lead to the conclusion that the local independence assumption is violated due to the corresponding subset of items.

Alternatively, one can model the local dependence structure. When taking this approach, it is important to impose certain parsimony/sparsity assumptions on the residual dependence structure to ensure the identifiability of the primary dimensions. Several modeling approaches have been taken. First, the local dependence may be modeled by additional latent traits with a sparse loading structure (e.g., Bradlow et al., 1999; Cai et al., 2011; Gibbons and Hedeker, 1992). These methods typically require knowing a priori the subsets of items, which are known as the testlets or item parcels, that require additional latent traits. Second, copula-based models have been proposed (Braeken, 2011; Braeken et al., 2007), where copula models are used for the conditional distribution of $(Y_1, ..., Y_J)$ given the primary latent dimensions. These models typically impose relatively strong parametric assumptions on the copula function. Finally, Markov random field IRT models have also been developed (Chen et al., 2018d; Hoskens and De Boeck, 1997; Ip, 2002a; Ip et al., 2004) that model the conditional distribution of $(Y_1, ..., Y_J)$ given the primary latent dimensions using a Markov random field model. Markov random field models (Koller and Friedman, 2009; Wainwright and Jordan, 2008), also known as undirected graphical models, are a powerful tool for modeling the dependence structure of multivariate data. Specifically, these models represent the conditional independence relationships between random variables using an undirected graph. When all the items are binary, this conditional model becomes an Ising model (Ising, 1925). Specifically, Chen et al. (2018d) proposed a flexible Markov random field IRT model and developed a Lasso-based method for learning a sparse Markov random field. Comparing with other methods, this approach does not require prior knowledge about the local dependence structure or impose a strong distributional assumption. It also gives a visualization of the local dependence structure, which facilitates the interpretation.

3.3.4 Item response functions. Finally, a key assumption of an IRT model is the functional form of the IRFs, especially when parametric IRFs are assumed. This assumption is commonly assessed by residual analysis. Hambleton et al.

(1991) first proposed item-specific standardized residuals for unidimensional IRT models that assess the fit of an IRF at different latent trait levels. Haberman and Sinharay (2013) and Haberman et al. (2013) modified the method of Hambleton et al. (1991) and further developed asymptotic normality results to provide theoretical guarantee for the residual-based statistical inference.

Alternatively, one may also fit IRT models with non-parametric IRFs and then compare the fitted parametric IRFs with their non-parametric counterparts. See Section 2.8 for a review of non-parametric IRT models.

3.4 Learning the Latent Structure

3.4.1 DIF analysis. DIF analysis, which is introduced in Section 2.8, is essential for ensuring the validity of educational and psychological measurement. The basic idea of DIF analysis is to compare multiple groups for their performance on one or multiple potential items, taking into consideration that the groups may have different ability distributions. The DIF problem can be viewed as a multi-group comparison problem. However, it differs from standard settings of the analysis of variance or covariance in that the latent trait values, which are not directly observed, need to be controlled in the DIF analysis. Many statistical procedures have been developed for DIF analysis; see Holland and Wainer (1993) and Roussos and Stout (2004) for a review.

Consider the DIF formulation described in Section 2.8 under a MIMIC model. Recall that $x_i = 0$ and 1 indicate the reference and focal group memberships, respectively. In the DIF analysis, the IRFs may depend on the group membership. That is, we define a group-specific IRF as $g_j(\boldsymbol{\theta}|\boldsymbol{\pi}_j^x) = P(Y_{ij} = 1|\boldsymbol{\theta}_i = \boldsymbol{\theta}, x_i = x)$, where $\boldsymbol{\pi}_j^0$ and $\boldsymbol{\pi}_j^1$ denote the group-specific item parameters. In addition, the distribution of $\boldsymbol{\theta}_i$ can depend on the group membership. Under this formulation, the DIF items are those satisfying $\boldsymbol{\pi}_j^0 \neq \boldsymbol{\pi}_j^1$. That is, two individuals from different groups would perform differently on a DIF item, even when they have the same latent trait levels. In educational testing, it could mean an item being more difficult for a test taker from one group than one from the other group even though the two test takers have the same ability level, which is a threat to the fairness of the test. Under the path diagram in panel (c) of Figure 2, the DIF items are the ones with directed edges from x. The goal of DIF analysis is to detect all the DIF items from data, i.e., to learn the edges from x to Y_js.

The DIF analysis is complicated by the unobserved $\boldsymbol{\theta}_i$, which needs to be conditioned upon in the group comparison. This complication is due to an intrinsic identifiability issue. More specifically, consider the MIMIC model example given in Section 2.8. The model remains unchanged if we subtract any constant c_0 from the parameter β in the marginal distribution of θ_i given x_i and further replace γ_j in the item response function by $\gamma_j + a_j c_0$. To ensure identifiability, one needs to assume that there are a sufficient number of DIF-free items to fix the latent variable distribution, which becomes assuming $\gamma_j = 0$ for many j in this specific MIMIC model.

The Mantel-Haenszel procedure (Holland and Thayer, 1988) is widely adopted for DIF analysis. This procedure compares the two groups' performance on an item based on the Mantel-Haenszel statistic (Mantel and Haenszel, 1959), using the total score as a matching variable. By stratifying data using the total score, which serves as a proxy of the latent ability being measured, the group difference

in the latent ability distributions is accounted for in the comparison. This procedure has a theoretical guarantee under the Rasch model. That is, the p-value given by the Mantel-Haenszel procedure is valid when data follow the Rasch model, for which the total score is a sufficient statistic for the person-specific latent variable. Note that the Mantel-Haenszel procedure implicitly assumes that the rest of the items are DIF-free to fix the latent variable distribution when analysing one item.

Another popular approach to DIF analysis takes a general nonparametric IRT setting, as first proposed in Shealy and Stout (1993a,b). Specifically, Shealy and Stout (1993a,b) laid a theoretical foundation for DIF analysis from a nonparametric multidimensional IRT perspective. Under this general statistical framework, they provided mathematical characterizations of different DIF types and further pointed out the importance of simultaneously analyzing the bias in a set of items. They also proposed a nonparametric procedure, known as the SIBTEST. Similar to the Mantel-Haenszel procedure, SIBTEST compares the overall performance of the reference and focal groups on one or multiple suspected items to detect DIF, by matching individuals on a known subset of DIF-free items. Further developments in this direction include Chang et al. (1996), Douglas et al. (1996), Jiang and Stout (1998), Li and Stout (1996), Nandakumar and Roussos (2004), Stout et al. (1997), among others.

Both the Mantel-Haenszel and SUBTEST procedures are hypothesis-testing-based. Possibly more naturally, one can view the DIF problem as a model selection problem. More specifically, in the specific MIMIC example discussed previously, the DIF problem becomes to find the non-zero γ_js, under the sparsity assumption described above. Taking this view, several regularized estimation methods have been developed for the detection of DIF items under MIMIC-type models; see Belzak and Bauer (2020), Huang (2018), Magis et al. (2015), and Tutz and Schauberger (2015). In these methods, the DIF-free items are identified by LASSO-type regularizations.

3.4.2 Exploratory analysis of latent structure. As pointed out in Sections 2.1 and 2.8, the Q-matrix structure plays an important role in multidimensional IRT models. Although domain experts usually have ideas about the latent traits underlying the items, it is also helpful to consider an exploratory analysis setting under which the meanings of the latent traits and the structure of the Q-matrix are unknown. Under this setting, we may learn the latent traits and the Q-matrix from data and use the learned structure to validate the current hypotheses or generate new hypotheses about data.

Analytic rotation methods, which are originally developed for exploratory factor analysis, can be used to solve this structural learning problem under several multidimensional IRT models, including the M2PL model described in Section 2.1 and its probit counterpart in Section 2.4. More specifically, consider an exploratory M2PL model, for which there are no zero constraints on the loading matrix A, and $\boldsymbol{\theta}_i$ follows a standard multivariate normal distribution. Let $\hat{A}$ be an MML estimate of the loading matrix A, which is not unique due to the rotational indeterminacy. More specifically, replacing $\hat{A}$ by $\hat{A}T$ and replacing the covariance matrix of $\boldsymbol{\theta}_i$ by $T^{-1}(T^{-1})^\top$ lead to the same marginal likelihood function value. Matrix T is called an oblique rotation matrix when the diagonal entries of $T^{-1}(T^{-1})^\top$ all take value one. In the special case when the matrix T is an orthonormal matrix, $T^{-1}(T^{-1})^\top$ becomes an identity matrix and the

matrix T is known as an orthogonal rotation matrix. A rotation method minimizes a loss function $H(\hat{A}T)$ with respect to an oblique or orthogonal rotation matrix T, and then uses $\hat{A}\hat{T}$ as the final estimate of the loading matrix where $\hat{T}$ denotes the corresponding minimizer. The loss function is designed to impose approximate sparsity on the resulting loading matrix $\hat{A}\hat{T}$, implicitly assuming that there is an unobserved sparse Q-matrix. Different analytic rotation methods have been proposed that consider different loss functions, including Jennrich (2004, 2006), Jennrich and Bentler (2011), Jennrich and Sampson (1966), Kaiser (1958), among others. A comprehensive review of rotation methods can be found in Browne (2001). Among these methods, we draw attention to a special case of the component loss functions proposed in Jennrich (2004, 2006). This loss function takes the form $H(A) = \|A\|_1 = \sum_{j=1}^{J} \sum_{k=1}^{K} |a_{jk}|$. As we discuss below, it is closely related to the regularized estimation approach to this problem. The analytic rotation methods have some limitations. First, these methods do not give a sparse estimate of the loading matrix, though they may resolve the rotational indeterminacy issue and lead to a consistent estimator of the true loading matrix (Jennrich, 2004, 2006). Consequently, obtaining the Q-matrix estimate and further, the interpretation of the latent traits are not straightforward. Second, these methods are only applicable to multidimensional IRT models in which the latent traits follow a multivariate normal distribution and the IRFs take a simple generalized linear model form.

Regularized estimation methods have also been proposed that do not suffer from the above limitations. Specifically, consider the M2PL model. A LASSO-based regularization method solves the optimization

$$(3.7) \quad (\hat{\boldsymbol{\pi}}_1^{\lambda}, ..., \hat{\boldsymbol{\pi}}_J^{\lambda}, \hat{\Sigma}^{\lambda}) = \underset{\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, \Sigma}{\arg\min} \left(-\log L_{ML}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, \Sigma) + \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} |a_{jk}| \right),$$
$$\text{s.t. } \sigma_{kk} = 1, k = 1, ..., K.$$

With slight abuse of notation, $L_{ML}(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_J, \Sigma)$ denotes the marginal likelihood function. Note that we assume $\boldsymbol{\theta}_i$ to follow a multivariate normal distribution with mean zero and covariance matrix Σ whose diagonal entries all take value one. The Lasso penalty can lead to a sparse estimate of the loading matrix A, which further gives an estimate of the Q-matrix. In (3.7), λ is a non-negative tuning parameter that can be chosen by information criteria or cross-validation. Note that this approach is closely related to the rotation approach. That is, when λ converges to zero, the solution in (3.7) will converge to the oblique rotation solution under the L_1 loss function mentioned above. We note that the LASSO penalty (Tibshirani, 1996) in (3.7) can be replaced by non-convex penalty functions, including the widely used SCAD and MCP penalties (Fan and Li, 2001; Zhang, 2010). Compared with the analytic rotation methods, this approach is more general and can be applied to a wide range of IRT models, including multidimensional IRT models with continuous latent traits (Sun et al., 2016) and DCMs (Chen et al., 2015; Xu and Shang, 2018). The optimization for these regularized estimators, such as (3.7), is more complicated than that for MML estimation when the latent dimension is high due to the non-smooth penalty term. To solve this optimization problem, Zhang and Chen (2020) proposed a stochastic approximation algorithm that combines a standard stochastic approximation algorithm with the proximal

method (Parikh et al., 2014) to handle non-smooth penalty terms.

Besides the above methods, several alternative methods have been developed for DCMs. In particular, Chiu (2013) and de la Torre and Chiu (2016) proposed iterative search algorithms for Q-matrix validation and refinement. Chen et al. (2018a) and Culpepper (2019) proposed full Bayes methods for Q-matrix learning.

3.4.3 Learning the latent dimension. Statistical methods have been developed for determining the latent dimension K under an exploratory analysis setting. Some of them only apply to the probit model described in Section 2.4.2. These methods arise from exploratory linear factor analysis, by making use of the connection between the probit model and the linear factor models. More precisely, consider the underlying variable formulation for the probit model given in Section 2.4.2. When the true latent dimension is K, then the true correlation matrix of the underlying variables can be decomposed as the sum of a rank-K matrix and a diagonal matrix. Consequently, the latent dimensionality can be determined based on the estimated correlation matrix, which, as discussed in Section 3.1.2, can be efficiently computed. These methods include eigenvalue thresholding Kaiser (1960), subjective search for eigengap based on scree plot Cattell (1966), parallel analysis (Horn, 1965), among others.

Besides, information criteria are commonly used for determining the latent dimension. Specifically, the Akaike information criterion (Akaike, 1974) and Bayesian information criterion (Schwarz et al., 1978), which are calculated based on the marginal likelihood, are widely used for comparing IRT models of different latent dimensions (Cohen and Cho, 2016). Unlike the exploratory-factor-analysis-based methods, these information criteria are applicable to all parametric IRT models that treat the latent variables as random. Treating the latent variables as fixed parameters, Chen and Li (2021) proposed an information criterion based on the joint likelihood. Their method guarantees consistent selection of the latent dimension under a high-dimensional setting where the sample N and the number of items J grow to infinity simultaneously, and the latent dimension may also diverge.

Under nonparametric IRT modelling, Stout (1987, 1990) introduced the concept of essential dimensionality, aiming at finding the dominant dimensions in the data while allowing for the presence of ignorable nuisance dimensions. Several methods for assessing and choosing the essential dimension were proposed; see Stout (1987), Stout (1990), Nandakumar and Stout (1993), and Zhang and Stout (1999). They established consistency results under a double asysmptotic regime where both N and J grow to infinity.

4. NEW CHALLENGES AND OPPORTUNITIES

4.1 Measurement in Big Data Era

In the previous discussions, the IRT framework is tailored to the analysis of nicely structured item response data from traditional educational and psychological assessments. This specialization is a limitation for the generalization and extension of IRT methods to measurement problems based on more complex human behavioral data, which are becoming increasingly prevalent in the big data era.

For example, with the wide use in daily teaching of digital devices and informa-

tion technology (e.g., tablets, computers, and online grading systems), students' item responses to all the homework/exam questions throughout their entire academic life can be recorded. In addition, students' learning process data may also become available, including the processes of reading, searching, and problem-solving (e.g., Zhang and Chang, 2016). Moreover, new measurement tools are being developed to measure people's skills that are difficult to assess by traditional paper-and-pencil tests, such as critical thinking, collaboration, and complex problem-solving (OECD, 2018). For example, complex problem-solving skills are currently assessed in large-scale assessments, including the PISA, Programme for the International Assessment of Adult Competencies (PIAAC) and the National Assessment of Education Progress (NAEP) using simulated tasks in computers. Data collected were not only test-takers' success or failure in solving such tasks, but also their entire problem-solving processes recorded by computer log files that may contain important information about their complex problem-solving ability. In psychology, researchers tend to study psychological constructs using not only item response data but also a much wider range of daily behavioral data including those from social media, facial expression, and writing and speech (e.g., Inkster et al., 2016; Nave et al., 2018; Wang and Kosinski, 2018).

As revealed by the above examples, human behavioral data of much higher volume, variety, and velocity are being collected nowadays. It is widely believed that rich information in such big data can substantially improve our understanding of human behavior and facilitate its prediction. The obtained insights and prediction, assisted by information technology, can lead to effective personalized interventions, such as personalized learning systems that suggest learning strategies adaptive to person-specific learning history and characteristics and personalized monitoring and intervention system for mental health patients.

On the one hand, such big data can still be viewed as people's responses to complex items in our study and daily life. Thus, the fundamental latent dimension view of IRT should still apply to such big data. It is reasonable to believe that people's behaviors are mainly driven by some intrinsic and relatively stable latent traits. Therefore, such latent traits may be measured and be used to predict people's future behaviors. On the other hand, compared with traditional item response data, big data have a far more complex structure and a much higher volume. As a result, most of the traditional IRT models and algorithms are no longer suitable.

The question now becomes how the methodological framework of IRT may be extended to embrace the opportunities and challenges in the big data era. While taking educational and psychological theories as the theoretical ground and ensuring the primacy of measurement validity, more flexible models and computationally faster algorithms from statistical and machine learning fields may be incorporated into the IRT framework to better measure human traits. The following sections will discuss two future directions that we believe are important and require methodological developments. We also note that many other promising directions are omitted here due to the space limit.

4.2 Learning the Structure of Fine-grained Constructs

The ultimate goal of psychological measurement is to understand human behavioral patterns, i.e., to obtain knowledge about how people learn, think, and

make decisions. Such knowledge is often gained from taxonomic analysis of individual differences. With big data, evidence-based taxonomy of educational or psychological constructs at a higher resolution becomes possible. For example, in the field of psychopathology, the traditional classification system of mental health disorders, widely adopted in the standard diagnostic criteria (*Diagnostic and Statistical Manual of Mental Disorders* (DSM; American Psychiatric Association, 2013) and the International Classification of Diseases (ICD; World Health Organization, 1992)), has received criticism for having limited reliability and validity (Kotov et al., 2017). A better-structured, finer-grained, and evidence-based taxonomic system needs to be developed (Kotov et al., 2017).

For another example, knowledge maps play an important role in designing a curriculum or in an e-learning system for conceptualizing and organizing the structure of the basic units of learning. They can be a powerful tool for guiding teaching and learning (Knight, 2013, Chapter 4). Each node of a knowledge map represents a concept or skill. An edge between two nodes captures their relationship, such as whether acquiring one concept requires another as a prerequisite. Examples can be found in Kingston and Broaddus (2017) on the development and use of digital knowledge maps for supporting formative assessments in mathematics, where a knowledge map can consist of thousands of nodes (points of mathematical knowledge) and thousands of connections between these nodes. Traditionally, knowledge maps were designed by domain experts, which are sometimes subjective. Rich data from daily teaching and e-learning platforms can be used to facilitate the design of knowledge maps. Data can tell us whether a node needs to be split into multiple ones and whether learning a certain concept/skill relies on certain concepts/skills as prerequisites.

These taxonomic problems are closely related to, but more complex than, the Q-matrix learning problem reviewed in Section 3.4.2. Specifically, data are more complex, multi-modal, and collected sequentially. For example, to design a knowledge map, one needs to consider the existing curriculum design, data from domain experts on each node's definition and the relationships between nodes, and students' learning history data. In addition, the improvement of the knowledge map is sequential, iterating between the steps of data collection under an existing knowledge map and updating the knowledge map. Moreover, a knowledge graph takes the form of a directed graph (Kingston and Broaddus, 2017), and the taxonomic structure of psychopathology is believed to be efficiently captured by a tree structure with many levels (Kotov et al., 2017). These structure learning problems are harder to formulate and solve than the previously discussed Q-matrix problem. Similar to the learning of Q-matrix in multidimensional IRT models (Chen et al., 2015; Liu et al., 2012, 2013; Xu and Shang, 2018), these taxonomic problems may be formulated under latent variable models and solved by similar regularized estimation methods for structure learning.

4.3 Measurement from a Prediction Perspective

Measurement and prediction are closely related. A major purpose of measurement is to make predictions, even for the traditional use of measurement results. For instance, an important use of test scores in education is to find the strengths and weaknesses of a student, which can provide predictions of the potential learning outcome in their subsequent learning trajectory. Accurate forecasting of stu-

dents' potential learning outcomes can help teachers to make a tailored teaching plan. In mental health, an individual's responses to a mental health questionnaire not only tell clinicians the individual's current mental health status but also reveal the likelihood of developing certain mental health disorders in the near future.

A related and more challenging problem is to make many predictions simultaneously. One such application is in personalized learning, where personalized formative assessment questions need to be selected from a large pool of items to meet the need of every student. A good formative assessment should achieve a balance in both its contents and difficulty/complexity, so that best serves the purpose of monitoring student learning. In this problem, to assemble a suitable set of formative assessment items, we need to predict every student's performance on each of the candidate items in the pool, which involves hundreds or even more simultaneous predictions for each student. Data in these predictive analyses may not be well-organized and may have massive missingness compared to conventional IRT analysis.

Latent variable models are a powerful tool that supports making such predictions because these models can learn the joint distribution of multivariate data even when massive missing data exist. One or multiple predictions can be made for the unobserved data given the observed ones with the learned joint distribution. Moreover, assumptions of latent variable models essentially reduce the dimensionality of the joint distribution of multivariate data, which will reduce the variance of prediction according to basic principles of statistical learning. Note that latent variable models, including nonlinear latent factor models and neural network models, have been popular tools for prediction tasks in other areas. These models may be borrowed to solve prediction problems in educational and psychological measurement. We note that some new developments have been made in this direction for analyzing and predicting log-file data of complex problem-solving processes; see Chen (2020), Tang et al. (2020), Tang et al. (2021), Wang et al. (2020), and Zhang et al. (2021).

Conventional IRT analysis takes a causal explanation perspective, which has substantial distinctions from the prediction perspective. While a model with high explanatory power may also have high predictive power, it is not always the case. It is well-known that a model selected for causal explanation may not have the best prediction performance (Shmueli, 2010). Therefore, a predictive model may take a very similar form as a multidimensional IRT model, estimating and selecting a predictive model often need to follow a different criterion from conventional IRT analysis. New criteria for estimation and model selection, such as cross-validation, need to be developed.

Learning a predictive model from data is computationally intensive, involving an iterative process that alternates between model fitting (solving an optimization problem) and model evaluation (via cross-validation). The computation becomes more demanding when many latent variables are needed to capture an individual's behavioral patterns. The problem is further exacerbated when online predictions need to be made (e.g., personalized learning). Given such high computational demand, it may be more suitable to use joint-likelihood-based estimators that treats the person-specific latent variables as fixed parameters (Chen et al., 2019, 2020a) for parameter estimation, due to its computational advantage.

Moreover, in some human behavior prediction problems, such as personalized learning, an individual's latent characteristics can substantially change within a relatively short period of time. Dynamic latent variable models need to be developed for solving such problems. We note that there have been some works in this direction, including Chen et al. (2018b) and Wang et al. (2018).

Finally, we point out that fairness is a key issue when making predictions, especially when interventions (e.g., learning interventions) are made based on the prediction results. The fairness issue is closely related to the DIF problem discussed in Section 3.4.1 that has also received wide attention in the statistics and machine learning communities (e.g., Barocas et al., 2018). Further developments that extend DIF analysis to ensure fairness are needed in the predictive modeling of multivariate behavioral data.

5. DISCUSSIONS

In this paper, we provide an overview of the statistical framework of IRT, and further showcase several applications in classical educational/psychological testing, such as test scoring, item calibration, computerized adaptive testing, differential item functioning, among others. Note that this list of topics is far from complete due to the space limit. Several important applications of IRT are not reviewed, including the detection of fraudulent behaviors and test assembly, among others. As we can see from the discussion, latent variables play a central role in IRT, serving as its philosophical foundation and methodological underpinning. We make the connection between the latent factor framework of IRT and new statistical techniques such as nonparametric methods, matrix completion, regularized estimation, and sequential analysis, and further offer a view of psychological measurement from the prediction perspective.

Thanks to the development of technology and data science, a personalization revolution is ongoing in all areas of our life, including mental health, education, social networking, among others (Lee, 2018). As a scientific discipline studying the measurement of individual characteristics supported by statistical methodology and substantive theory from psychology and education, psychometrics will play a key role in this personalization revolution. It will contribute to a better society, for example, by suggesting better learning strategies for learners and by providing early warnings to mental health patients through more accurate measurement and measurement supported prediction (Kapczinski et al., 2019; Zhang and Chang, 2016). As previously discussed, these problems need to be solved from two different perspectives – high-resolution taxonomic analysis for better explaining individual differences and predictive analytics for making personalized recommendations. Latent variable models and the previously discussed computationally efficient methods for the structure learning and estimation of latent variable models may play an important role in solving these new problems.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders,fifth edition, (DSM-5®)*. American Psychiatric Association, Arlington, VA.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32:283–301.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, 42:69–81.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley, Hoboken, NJ.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In Neyman, J., editor, *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150, Berkerly, CA. University of California Press.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, 75:292–308.

Bafumi, J., Gelman, A., Park, D. K., and Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, 13:171–187.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *Annals of Statistics*, 40:436–465.

Balsis, S., Ruchensky, J. R., and Busch, A. J. (2017). Item response theory applications in personality disorder research. *Personality Disorders: Theory, Research, and Treatment*, 8:298–308.

Barocas, S., Hardt, M., and Narayanan, A. (2018). *Fairness and machine learning*. fairmlbook.org. http://www.fairmlbook.org.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42:293–312.

Bartholomew, D. J. (1984). The foundations of factor analysis. *Biometrika*, 71:221–232.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, New York, NY.

Bartholomew, D. J. and Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology*, 55:1–15.

Bartholomew, D. J. and Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research*, 27:525–546.

Bartroff, J., Lai, T. L., and Shih, M.-C. (2012). *Sequential experimentation in clinical trials: Design and analysis*. Springer, New York, NY.

Belzak, W. and Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, 25:673–690.

Berridge, D. M. and Crouchley, R. (2011). *Multivariate generalized linear mixed models using R*. Chapman and Hall/CRC, Boca Raton, FL.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36:192–225.

Bhaskar, S. A. (2016). Probabilistic low-rank matrix completion from quantized measurements. *The Journal of Machine Learning Research*, 17:2131–2164.

Bhaskar, S. A. and Javanmard, A. (2015). 1-bit matrix completion under exact low-rank constraint. In *2015 49th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Birdsall, M. (2011). *Implementing computer adaptive testing to improve achievement opportunities*. Office of Qualifications and Examinations Regulation, Coventry, UK.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical theories of mental test scores*, pages 397–479. Addison-Wesley, Reading, MA.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37:29–51.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46:443–459.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, New York, NY.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64:153–168.

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. *Psychometrika*, 76:57–76.

Braeken, J., Tuerlinckx, F., and De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, 72:393–411.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36:111–150.

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

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35:307–335.

Cai, L., Yang, J. S., and Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16:221–248.

Cai, T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14:3619–3647.

Carlson, J. E. and von Davier, M. (2017). Item response theory. In Bennett, R. E. and von Davier, M., editors, *Advancing human assessment: A synthesis over seven decades*, pages 133–178. Springer, New York, NY.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1:245–276.

Cella, D., Chang, C.-H., and Heinemann, A. W. (2002). Item response theory (IRT): Applications in quality of life measurement, analysis and interpretation. In Mesbah, M., Cole, B. F., and Lee, M.-L. T., editors, *Statistical methods for quality of life studies*, pages 169–185. Springer, New York, NY.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., and Rose, M. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an nih roadmap cooperative group during its first two years. *Medical care*, 45:S3–S11.

Cen, H., Koedinger, K., and Junker, B. (2006). Learning factors analysis–a general method for cognitive model evaluation and improvement. In Ikeda, M., Ashley, K. D., and Chan, T.-W., editors, *International Conference on Intelligent Tutoring Systems*, pages 164–175. Springer.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80:1–20.

Chang, H.-H., Mazzeo, J., and Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, 33:333–353.

Chang, H.-H. and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20:213–229.

Chang, H.-H. and Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23:211–222.

Chang, H.-H. and Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37:1466–1488.

Chang, Y.-C. I. (2005). Application of sequential interval estimation to adaptive mastery testing. *Psychometrika*, 70:685–713.

Chen, W.-H. and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22:265–289.

Chen, Y. (2020). A continuous-time dynamic choice measurement model for problem-solving process data. *Psychometrika*, 85:1052–1075.

Chen, Y., Culpepper, S. A., Chen, Y., and Douglas, J. (2018a). Bayesian estimation of the dina q matrix. *Psychometrika*, 83:89–108.

Chen, Y., Culpepper, S. A., Wang, S., and Douglas, J. (2018b). A hidden markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement*, 42:5–23.

Chen, Y., Li, C., and Xu, G. (2021). A note on statistical inference for noisy incomplete 1-bit matrix. *arXiv preprint arXiv:2105.01769*.

Chen, Y. and Li, X. (2021). Determining the number of factors in high-dimensional generalised latent factor models. *Biometrika*. To appear.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2018c). Recommendation system for adaptive learning. *Applied Psychological Measurement*, 42:24–41.

Chen, Y., Li, X., Liu, J., and Ying, Z. (2018d). Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, 83:538–562.

Chen, Y., Li, X., and Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84:124–146.

Chen, Y., Li, X., and Zhang, S. (2020a). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115:1756–1770.

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, 110:850–866.

Chen, Y., Moustaki, I., and Zhang, H. (2020b). A note on likelihood ratio tests for models with latent variables. *Psychometrika*, 85:996–1012.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74:619–632.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25:573–578.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30:755–770.

Chernoff, H. (1972). *Sequential analysis and optimal design*. SIAM, Philadelphia, PA.

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37:598–618.

Chiu, C.-Y., Douglas, J. A., and Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74:633–665.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40:5–32.

Chung, K. L. (1954). On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25:463–483.

Cohen, A. S. and Cho, S.-J. (2016). Information criteria. In van der Linden, W., editor, *Handbook of item response theory: Models, statistical tools, and applications*. Chapman & Hall/CRC Press, Boca Raton, FL.

Collins, L. M., Fidler, P. L., Wugalter, S. E., and Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, 28:375–389.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91:729–737.

Cressie, N. and Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, 48:129–141.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, 40:454–476.

Culpepper, S. A. (2019). Estimating the cognitive diagnosis q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika*, 84:333–357.

Davenport, M. A., Plan, Y., van den Berg, E., and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference*, 3:189–223.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7:1–46.

De Boeck, P. and Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer, New York, NY.

De Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., and Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, 45:104–115.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76:179–199.

de la Torre, J. and Chiu, C.-Y. (2016). A general method of empirical q-matrix validation. *Psychometrika*, 81:253–273.

de Leeuw, J. and Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, 11:183–196.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In Nichols, P. D., Chipman, S., and Brennan, R. L., editors, *Cognitively diagnostic assessment*, pages 361–389. Routledge, New York, NY.

Diebolt, J. and Ip, E. H. (1996). Stochastic EM: Method and application. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov chain Monte Carlo in practice*, pages 259–273. Springer, New York, NY.

Doignon, J.-P. and Falmagne, J.-C. (2012). *Knowledge spaces*. Springer, New York, NY.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability

estimation. *Psychometrika*, 62:7–28.

Douglas, J. A. (2001). Asymptotic identifiability of nonparametric item response models. *Psychometrika*, 66:531–540.

Douglas, J. A., Stout, W., and DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics*, 21:333–363.

Drton, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3):474–497.

Efron, B. (2003). Robbins, empirical Bayes and microarrays. *The Annals of Statistics*, 31:366–378.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68:117–130.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96:1348–1360.

Fang, G., Liu, J., and Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika*, 84:19–40.

Fang, G., Xu, X., Guo, J., Ying, Z., and Zhang, S. (2021). Identifiability of bifactor models. *Statistica Sinica*. In press.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, 7:19–29.

Feuerverger, A., He, Y., and Khatri, S. (2012). Statistical significance of the Netflix challenge. *Statistical Science*, 27:202–231.

Fienberg, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments. *Journal of the American Statistical Association*, 95:643–647.

Fienberg, S. E. and Meyer, M. M. (1983). Loglinear models and categorical data analysis with psychometric and econometric applications. *Journal of Econometrics*, 22:191–214.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, 23:165–170.

Gibbons, R. D. and Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57:423–436.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*, 40:979–1001.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning.* MIT Press, Cambridge, MA.

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach. *American Journal of Sociology*, 79:1179–1259.

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences*, 95:7270–7274.

Gu, Y. and Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*, 48:2082–2107.

Guo, H. and Sinharay, S. (2011). Nonparametric item response curve estimation with correction for measurement error. *Journal of Educational and Behavioral Statistics*, 36:755–778.

Haberman, S. J. (1977a). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 5:815–841.

Haberman, S. J. (1977b). Product models for frequency tables involving indirect observation. *The Annals of Statistics*, 5:1124–1147.

Haberman, S. J. (2005). Identifiability of parameters in item response models with unconstrained ability distributions (ETS Research Report Series RR-05-24). Technical report, Educational Testing Service, Princeton, NJ.

Haberman, S. J. and Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of the American Statistical Association*, 108:1435–1444.

Haberman, S. J., Sinharay, S., and Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, 78:417–440.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26:301–321.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage, Newbury Park, CA.

Hays, R. D., Morales, L. S., and Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38:II28–II42.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74:191–210.

Holland, P. W. (1990a). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, 55:5–18.

Holland, P. W. (1990b). On the sampling theory foundations of item response theory models. *Psychometrika*, 55:577–601.

Holland, P. W. and Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68:123–149.

Holland, P. W. and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H. and Braun, H. I., editors, *Test validity*, pages 129–145. Lawrence Erlbaum Associates, Hillsdale, NJ.

Holland, P. W. and Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30:179–185.

Hoskens, M. and De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, 2:261 – 275.

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, 71:499–522.

Huber, P., Ronchetti, E., and Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:893–908.

Inkster, B., Stillwell, D., Kosinski, M., and Jones, P. (2016). A decade into Facebook: Where is psychiatry in the digital age? *The Lancet Psychiatry*, 3:1087–1090.

Ip, E. H. (2002a). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, 67:367–386.

Ip, E. H. (2002b). On single versus multiple imputation for a class of stochastic algorithms estimating maximum likelihood. *Computational Statistics*, 17:517–524.

Ip, E. H., Wang, Y. J., De Boeck, P., and Meulders, M. (2004). Locally dependent latent trait model for polytomous responses with application to inventory of hostility. *Psychometrika*, 69:191–216.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31:253–258.

Jennrich, R. I. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika*, 69:257–273.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, 71:173–191.

Jennrich, R. I. and Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, 76:537–549.

Jennrich, R. I. and Sampson, P. (1966). Rotation for simple loadings. *Psychometrika*, 31(3):313–323.

Jiang, H. and Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics*, 23:291–322.

Jiang, Z. and Templin, J. (2019). Gibbs samplers for logistic item response models via the Pólya–Gamma distribution: A computationally efficient data-augmentation strategy. *Psychometrika*, 84:358–374.

Johnson, M. S. (2006). Nonparametric estimation of item and respondent locations from unfolding-type items. *Psychometrika*, 71:257–279.

Johnson, M. S. (2007). Modeling dichotomous item responses with free-knot splines. *Computa-*

tional Statistics & Data Analysis, 51:4178–4192.

Jöreskog, K. G. and Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36:347–387.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25:258–272.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20:141–151.

Kapczinski, F., Mwangi, B., and Passos, I. C., editors (2019). *Personalized psychiatry: Big data analytics in mental health*. Springer, New York, NY.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., and Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, 56:4243–4258.

Khajah, M., Wing, R., Lindsey, R., and Mozer, M. (2014a). Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In Stamper, J., Pardos, Z., Mavrikis, M., and McLaren, B., editors, *Proceedings of the 7th International Conference on Educational Data Mining*, pages 99–106, London, United Kingdom.

Khajah, M. M., Huang, Y., González-Brenes, J. P., Mozer, M. C., and Brusilovsky, P. (2014b). Integrating knowledge tracing and item response theory: A tale of two frameworks. In Kravcik, M., Santos, O., and Boticario, J., editors, *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE 2014) at the 22th international conference on user modeling, adaptation, and personalization*, pages 7–12, Aalborg, Denmark.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 27:887–906.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, 77:153–162.

Kingston, N. and Broaddus, A. (2017). The use of learning map systems to support the formative assessment in mathematics. *Education Sciences*, 7:41.

Knight, J. (2013). *High-impact instruction: A framework for great teaching*. Corwin Press, Thousand Oaks, CA.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT Press, Cambridge, MA.

Koren, Y. and Bell, R. (2015). Advances in collaborative filtering. *Recommender systems handbook*, pages 77–118.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., et al. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, 126:454–477.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, 11:303–351.

Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *The Annals of Statistics*, 7:1196–1221.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 61:273–287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, 62:74–82.

Lazarsfeld, P. and Henry, N. (1968). *Latent structure analysis*. Houghton Mifflin, New York, NY.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In Stouffer, S. A., editor, *Studies in social psychology in World War II, Vol. 4: Measurement and prediction*, pages 362–412. Princeton University Press, Princeton, NJ.

Lazarsfeld, P. F. (1959). Latent structure analysis. *Psychology: A study of a Science*, 3:476–543.

Lee, K.-F. (2018). *AI superpowers: China, Silicon Valley, and the new world order*. Houghton Mifflin Harcourt, New York, NY.

Lee, S.-Y., Poon, W.-Y., and Bentler, P. M. (1990). A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika*, 55:45–51.

Lee, S.-Y., Poon, W.-Y., and Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, 57:89–105.

Lee, S.-Y., Poon, W.-Y., and Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, 48:339–358.

Li, H.-H. and Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, 61:647–677.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86:96–107.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics*, 80:221–239.

Liu, J., Xu, G., and Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, 36:548–564.

Liu, J., Xu, G., and Ying, Z. (2013). Theory of the self-learning Q-matrix. *Bernoulli*, 19:1790 – 1817.

Liu, J., Ying, Z., and Zhang, S. (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika*, 80:468–490.

Liu, Y. and Maydeu-Olivares, A. (2013). Local dependence diagnostics in irt modeling of binary data. *Educational and Psychological Measurement*, 73:254–274.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28:989–1020.

Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, 31:3–31.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1:95–100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Routledge, New York, NY.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley, Reading, MA.

Macready, G. B. and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2:99–120.

Magis, D., Tuerlinckx, F., and De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, 40:111–135.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47:149–174.

Maydeu-Olivares, A. and Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables: a unified framework. *Journal of the American Statistical Association*, 100:1009–1020.

Maydeu-Olivares, A. and Joe, H. (2006). Limited information goodness-of-fit testing in multi-dimensional contingency tables. *Psychometrika*, 71:713 – 732.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* Chapman and Hall/CRC, London, UK.

Meng, X.-L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, 91:1254–1267.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49:359–381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80:993–997.

Mislevy, R. J. and Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, 13:57–75.

Mislevy, R. J. and Wu, P.-K. (1996). Missing responses and irt ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, 1996:i–36.

Mokkan, R. J. and Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6:417–430.

Mokken, R. (1971). *A theory and procedure of scale analysis.* Mouton, The Hague, The Netherlands.

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65:391–411.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied*

Psychological Measurement, 16:159–176.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, 43:551–560.

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

Muthén, B. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, 10:121–132.

Muthén, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In Howard, W. and Braun, H. I., editors, *Test validity*, pages 213–238. Routledge, New York, NY.

Nandakumar, R. and Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics*, 29:177–199.

Nandakumar, R. and Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, 18:41–68.

Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., and Rentfrow, J. (2018). Musical preferences predict personality: Evidence from active listening and facebook likes. *Psychological Science*, 29:1145–1158.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3:1–18.

OECD (2018). *The future of education and skills: Education 2030*. Organisation for Economic Co-operation and Development (OECD), Paris, France.

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, 1:127–239.

Pearson, K. (1900a). I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195:1–47.

Pearson, K. (1900b). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855.

Puterman, M. L. (2014). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Hoboken, NJ.

Rabe-Hesketh, S. and Skrondal, A. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC, New York, NY.

Ramsay, J. and Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, 56:365–379.

Ramsay, J. O. and Abrahamowicz, M. (1989). Binomial regression with monotone splines: A psychometric application. *Journal of the American Statistical Association*, 84:906–915.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Nielsen and Lydiche, Copenhagen, Denmark.

Reckase, M. (2009). *Multidimensional item response theory*. Springer, New York, NY.

Reise, S. P. and Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, 5:27–48.

Reiser, M. (1996). Analysis of residuals for the multionmial item response model. *Psychometrika*, 61:509–528.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1:33–49.

Robbins, H. (1956). An empirical Bayes approach to statistics. In Neyman, J., editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163, Berkeley, CA. University of California Press.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407.

Robin, F., Steffen, M., and Liang, L. (2014). The multistage test implementation of the GRE revised General Test. In Yan, D., von Davier, A. A., and Lewis, C., editors, *Computerized Multistage Testing: Theory and Applications*, pages 325–341. Chapman and Hall/CRC, Boca

Raton, FL.

Robinson, P. (1974). Identification, estimation and large-sample theory for regressions containing unobservable variables. *International Economic Review*, 15:680–692.

Roussos, L. A. and Stout, W. (2004). Differential item functioning analysis. In Kaplan, D., editor, *The Sage handbook of quantitative methodology for the social sciences*, pages 107–116. Sage, Thousand Oaks, CA.

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, New York, NY.

Ruppert, D. (1988). Efficient estimators from a slowly convergent Robbins-Monro procedure. *School of Oper. Res. and Ind. Eng., Cornell Univ., Ithaca, NY, Tech. Rep No. 781*.

Rutkowski, L., von Davier, M., and Rutkowski, D. (2013). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. CRC Press, Boca Raton, FL.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17:1–68.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Searle, S. R. and McCulloch, C. E. (2001). *Generalized, linear, and mixed models*. John Wiley & Sons, New York, NY.

Shealy, R. and Stout, W. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58:159–194.

Shealy, R. T. and Stout, W. F. (1993b). An item response theory model for test bias and differential test functioning. In Holland, P. and Wainer, H., editors, *Differential test functioning*, pages 197–239. Lawrence Erlbaum Associates, Hillsdale, NJ.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25:289–310.

Sijtsma, K. and Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage, Thousand Oaks, CA.

Sijtsma, K. and van der Ark, L. A. (2017). A tutorial on how to do a mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, 70:137–158.

Spearman, C. (1904a). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, 15:201–292.

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72–101.

Stout, W., Li, H.-H., Nandakumar, R., and Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement*, 21:195–213.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52:589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55:293–325.

Sun, J., Chen, Y., Liu, J., Ying, Z., and Xin, T. (2016). Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika*, 81:921–939.

Tang, X., Chen, Y., Li, X., Liu, J., and Ying, Z. (2019). A reinforcement learning approach to personalized learning recommendation systems. *British Journal of Mathematical and Statistical Psychology*, 72:108–135.

Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, 85:378–397.

Tang, X., Wang, Z., Liu, J., and Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, 74:1–33.

Tatsuoka, C. and Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:143–157.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20:345–354.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11:287–305.

Thissen, D. and Steinberg, L. (2009). Item response theory. In Milsap, editor, *The Sage handbook*

of quantitative methods in psychology, pages 148–177.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, 9:23–30.

Tutz, G. and Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, 80:21–43.

van der Linden, W. J. (2018). *Handbook of item response theory*. Chapman and Hall/CRC, New York, NY.

van der Linden, W. J. and Glas, C. A. (2010). *Elements of adaptive testing*. Springer, New York, NY.

Vasdekis, V. G., Cagnone, S., and Moustaki, I. (2012). A composite likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika*, 77:425–441.

Vasdekis, V. G., Rizopoulos, D., and Moustaki, I. (2014). Weighted pairwise likelihood estimation for a general class of random effects models. *Biostatistics*, 15:677–689.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61:287–307.

von Davier, M. and Lee, Y.-S., editors (2019). *Handbook of Diagnostic Classification Models*. Springer, New York, NY.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge, New York, NY.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1:1–305.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16:117–186.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19:326–339.

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43:57–87.

Wang, Y. and Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114:246–257.

Wang, Z., Tang, X., Liu, J., and Ying, Z. (2020). Subtask analysis of process data through a predictive model. *arXiv preprint arXiv:2009.00717*.

Wedin, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111.

Weiss, D. J. (1974). Strategies of adaptive ability measurement (research report 74–5). Technical report, University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis, MN.

Weiss, D. J., editor (1978). *Proceedings of the 1977 Computerized Adaptive Testing Conference*. University of Minnesota, Department of Psychology, Psychometric Methods Program, Minneapolis, MN.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6:473–492.

Weiss, D. J. and Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21:361–375.

Wilson, K. and Nichols, Z. (2015). The Knewton platform: A general-purpose adaptive learning infrastructure. White paper, Knewton, New York, NY.

Winsberg, S., Thissen, D., and Wainer, H. (1984). Fitting item characteristic curves with spline functions. *ETS Research Report Series*, 84:1–14.

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

Wood, R. L., Wingersky, M. S., and Lord, F. M. (1978). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters*. Educational Testing Service.

World Health Organization (1992). *The ICD-10 classification of mental and behavioural disorders*. World Health Organization, Geneva, Switzerland.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals*

of Statistics, 45:675–707.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113:1284–1295.

Xu, G. and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, 81:625–649.

Xu, X., Chang, H., and Douglas, J. (2003). A simulation study to compare cat strategies for cognitive diagnosis. In *Annual meeting of the American Educational Research Association*, Chicago, IL.

Yan, D., Von Davier, A. A., and Lewis, C. (2016). *Computerized multistage testing: Theory and applications*. CRC Press, Boca Raton, FL.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8:125–145.

Zellner, A. (1970). Estimation of regression relationships containing unobservable independent variables. *International Economic Review*, 11:441–454.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Annals of Statistics*, 31:379–390.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38:894–942.

Zhang, J. and Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64:213–249.

Zhang, S. and Chang, H.-H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, 1:67–92.

Zhang, S. and Chen, Y. (2020). Computation for latent variable model estimation: A unified stochastic proximal framework. *arXiv preprint arXiv:2008.07214*.

Zhang, S., Chen, Y., and Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology*, 73:44–71.

Zhang, S., Wang, Z., Qi, J., Liu, J., and Ying, Z. (2021). Accurate assessment via process data. *arXiv preprint arXiv:2103.15034*.

Zhu, Y., Shen, X., and Ye, C. (2016). Personalized prediction and sparsity pursuit in latent factor models. *Journal of the American Statistical Association*, 111:241–252.