Journal of Educational and Behavioral Statistics
Fall 1998, Vol. 23, No. 3, pp. 236–243

Item Response Theory Models Applied to Data Allowing Examinee Choice

Eric T. Bradlow
University of Pennsylvania

Neal Thomas
University of North Carolina

Keywords: *examinee choice, item response theory, missing data*

Examinations that permit students to choose a subset of the items are popular despite the potential that students may take examinations of varying difficulty as a result of their choices. We provide a set of conditions for the validity of inference for Item Response Theory (IRT) models applied to data collected from choice-based examinations. Valid likelihood and Bayesian inference using standard estimation methods require (except in extraordinary circumstances) that there is no dependence, after conditioning on the observed item responses, between the examinees choices and their (potential but unobserved) responses to omitted items, as well as their latent abilities. These independence assumptions are typical of those required in much more general settings. Common low-dimensional IRT models estimated by standard methods, though potentially useful tools for educational data, do not resolve the difficult problems posed by choice-based data.

Introduction

In many currently-used assessments of educational ability (e.g., Advanced Placement examinations), examinees are instructed to respond to a subset of items from a designated list. One purpose of such test designs is to give each examinee the opportunity to answer items on a wide range of topics, yet limit the time the assessment takes to complete (hence only a subset of the items is answered by each examinee). Choice of items is also popular because it is perceived as a more realistic, or authentic task. In such paradigms, missing data (the responses to those items not chosen) exist by design. Estimation of examinee ability in this choice setting can thus be examined in the missing data framework proposed by Little and Rubin (1987). Allen, Holland, and Thayer (1994); Wainer and Thissen (1994); Wainer, Wang, and Thissen (1994); and

The authors wish to thank Robert J. Mislevy, Educational Testing Service for comments on an early version of the paper, Minhwei Wang, Educational Testing Service, for computation support, and the AE and reviewers for suggestions which helped strengthen the paper.

Wang, Wainer, and Thissen (1995) discuss this missing data perspective applied to choice-based exams. Under this framework, described in the next section, it is shown that only in very special cases can valid inferences about examinee abilities be obtained using common methods that ignore the choice mechanism (i.e., ignoring the fact that the examinees choose the items to which they respond).

A Missing Data Formulation

We utilize the notation given in Little and Rubin (1987, p. 89–90) to describe a likelihood approach to inference for examinee abilities in an assessment with self-selected observed scores. The description of probability models, data, parameters, inferential goals, and intuition is given using language associated with standard IRT models. It is understood however, that these results have much wider applicability.

Let $\mathbf{Y}$ denote the vector of scores for an examinee in the absence of missing data (i.e., if the examinee answered all of the items). We further define $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$, where $\mathbf{Y}_{obs}$ denotes the observed scores and $\mathbf{Y}_{mis}$ the missing scores, $R = (r_i)$ a response indicator with $r_i = 1$ if the score of item i is observed and 0 otherwise, θ an examinee ability parameter, which can be multidimensional, and β a set of parameters describing test item properties (e.g., item difficulties).

Inferences for θ and β in the standard IRT setting (without choice) are derived from the product of the observed score likelihood functions, $f(\mathbf{Y}_{obs}|\theta,\beta)$, for each examinee. When allowing examinee choice, the observed data are extended to include the observed scores, Y_{obs}, and the choice indicators, R. The contribution of $\mathbf{Y}_{obs}$ and R from an examinee to the likelihood is

$$f[(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}), R|\theta,\beta,\psi] = f[R|(\mathbf{Y}_{obs},\mathbf{Y}_{mis}),\theta,\beta,\Psi]\, f[(\mathbf{Y}_{obs},\mathbf{Y}_{mis})|\theta,\beta], \quad (1)$$

where Ψ is a set of parameters associated with the conditional distribution of R given the data $\mathbf{Y}$, the examinee ability, and the IRT parameters. When different forms of tests are randomly assigned to students, and no choice is allowed, R does not depend on $\mathbf{Y}$, θ, or β, and the parameters in Ψ are the proportions of examinees assigned to each item. An example of a more complex choice model assumes that the distribution of the R is a polychotomous logistic regression depending on the ability, θ, and the parameters in Ψ are the coefficients of these regressions. Equation (1) is a special case of the factorization of $f[(\mathbf{Y}_{obs},\mathbf{Y}_{mis}),R|\theta,\beta,\Psi]$ described in Little and Rubin (1987, p. 89).

Valid inferences for θ and β allowing choice can be obtained from the standard IRT likelihood function $f(\mathbf{Y}_{obs}|\theta,\beta)$ (which ignores choice) in general only if $f(\mathbf{Y}_{obs},R|\theta,\beta,\Psi) \propto f(\mathbf{Y}_{obs}|\theta,\beta)$ as a function of θ and β. We consider assumptions sufficient to ensure this condition. Using (1) we write

$$f(\mathbf{Y}_{obs},R|\theta,\beta,\Psi) = \int f[(\mathbf{Y}_{obs},\mathbf{Y}_{mis}),R|\theta,\beta,\Psi]d\mathbf{Y}_{mis}$$
$$= \int f[R|(\mathbf{Y}_{obs},\mathbf{Y}_{mis}),\theta,\beta,\Psi]f[(\mathbf{Y}_{obs},\mathbf{Y}_{mis})|\theta,\beta]d\mathbf{Y}_{mis} \quad (2)$$

Under this factorization, two conditions that permit inference based on $f(\mathbf{Y}_{obs}|\theta,\beta)$ are: (a) $f[R|(\mathbf{Y}_{obs},\mathbf{Y}_{mis}),\theta,\beta,\Psi] = f(R|\mathbf{Y}_{obs},\theta,\beta,\Psi)$, known as the missing at random (MAR) assumption; and (b) $f(R|\mathbf{Y}_{obs},\theta,\beta,\Psi) = f(R|\mathbf{Y}_{obs},\Psi)$, with Ψ distinct from θ, and β, the assumption of ignorability of the missing data mechanism (Rubin, 1987, p. 53). With assumptions (a) and (b), $f(\mathbf{Y}_{obs},R|\theta,\beta,\Psi) \propto f(\mathbf{Y}_{obs}|\theta,\beta)$, the usual IRT likelihood function, and standard maximum likelihood and Bayesian estimation methods are applicable.

Assumption (a) implies that examinees cannot identify items that would ordinarily be difficult for a student of their ability, but are manageable for them because, for example, they had recent classroom exposure to a similar item. If the MAR assumption is true, it eliminates one of the primary motivations for allowing choice. Assumption (b) is more easily overlooked (e.g., Wainer & Thissen, 1994; Wainer, Wang, & Thissen, 1994, 1995); students of differing ability must not select broadly easier (or difficult) items. This could happen if, for example, weaker students were attracted to items involving more basic concepts, even though the items might be very difficult. Mislevy and Sheehan (1989) provide an example of invalid inference that can result from standard IRT analyses when students of differing ability are assigned different items.

Note that condition (b) is made more plausible by the fact that choice and ability are required to be independent only after conditioning on the observed item responses. Even with a long test, invalid inference can result if there are a substantial number of choice items, as demonstrated in Section 3. We can be confident of the approximate validity of assumptions (a) and (b) without substantial experimentation only when the choice component of an examination is so small that it cannot meaningfully change the overall scores assigned.

A Simulated Example

To demonstrate the potential for bias in parameter estimates using standard IRT methods under violations of assumptions (a) and (b), we conducted three simulation experiments using 200 randomly generated normal $\phi(0, 1)$ item difficulties, β, and 5000 randomly generated $\phi(0, 1)$ examinee abilities, θ, to construct a 5000 $\times$ 200-dimensional matrix of Rasch model binary responses. Item parameter estimates were obtained using a marginal maximum likelihood procedure (Bock & Aitkin, 1981) with the mean of the distribution constrained to be zero; ability parameter estimates were computed using maximum likelihood with the item parameters fixed at their estimated values.

In each of the three experiments, the 5000 examinees answered the first 50 items. This corresponds to a mandatory section on each test form. The remaining items were partitioned into 75 pairs of items.

In the first experiment, an item within each of the 75 pairs was *randomly* assigned to each examinee. This design, which satisfies assumptions (a) and (b), is used as a baseline for comparison. It results in valid inferences for both the examinee abilities and the item parameters as shown in Figures 1a and 2a. In Figure 1a, the difference between the estimated and true item parameter is plotted against the true item parameter. The corresponding plot for examinee abilities is shown in Figure 2a. We identify the parameter estimates for the 50 mandatory items with an "x", and the 75 choice pairs with an "o". No systematic pattern exists in the estimated differences and each point cloud is centered at zero.

The second and third simulation experiments were designed to violate the MAR assumption and ignorability assumptions given in (a) and (b), respectively. We used exaggerated choice mechanisms to produce very large and easy-to-detect biases. Very little is currently known about the nature and magnitude of realistic violations of the ignorability and MAR assumptions. We hope to stimulate more interest in experiments which yield useful data for modeling the missing indicators R.

Within each pair for the second experiment, if the responses were either both 0 or both 1, then the examinee chose the first item with probability Ψ_1, and the second item with probability $1 - \Psi_1$. If the responses within a pair were (0,1) or (1,0), then the examinee chose the correct one with probability Ψ_2, and the incorrect one with probability $1 - \Psi_2$. We simulated the exams using $\Psi_1 = .5$, and $\Psi_2 = .75$. This data set violates the MAR assumption because the $f(R|(\mathbf{Y}_{obs},\mathbf{Y}_{mis}),\theta,\beta,\Psi)$ depends on the value of $\mathbf{Y}_{mis}$: specifically, the items the examinee selected were more likely to be correct.

In the third experiment, items were selected within pairs based on the difficulty of the two items determined by their item parameters in β, and the ability of the examinee. Specifically, the better examinees (defined as those with ability greater than zero) chose the easier of the two items in the pair with probability Ψ_1, whereas the weaker examinees (ability less than zero) chose the easier item with probability Ψ_2, and the harder item with probability $1 - \Psi_2$. We simulated the exams with $\Psi_1 = .95$, and $\Psi_2 = .5$. This describes a process where the better examinees look over both items in a pair, and can often decide which of the two is easier. The choice mechanism $f(R|(\mathbf{Y}_{obs},\mathbf{Y}_{mis}),\theta,\beta,\Psi)$ depends upon θ,β, and Ψ, and therefore is not ignorable. A choice mechanism where the better examinees can more aptly choose items may not hold in practice (e.g., Wang, 1992).

The differences in the estimated and true item parameters plotted against the true item parameters for the second and third experiments are given in Figures 1b and 1c, respectively. The corresponding plots for examinee abilities are given in Figures 2b and 2c.

For experiment 2, Figure 1b indicates a consistent underestimation of item difficulty for the 50 mandatory items and a more severe underestimation bias for the remaining 75 choice pairs. We underestimated the choice item difficulties as anticipated, because students tended to select them only when they knew the

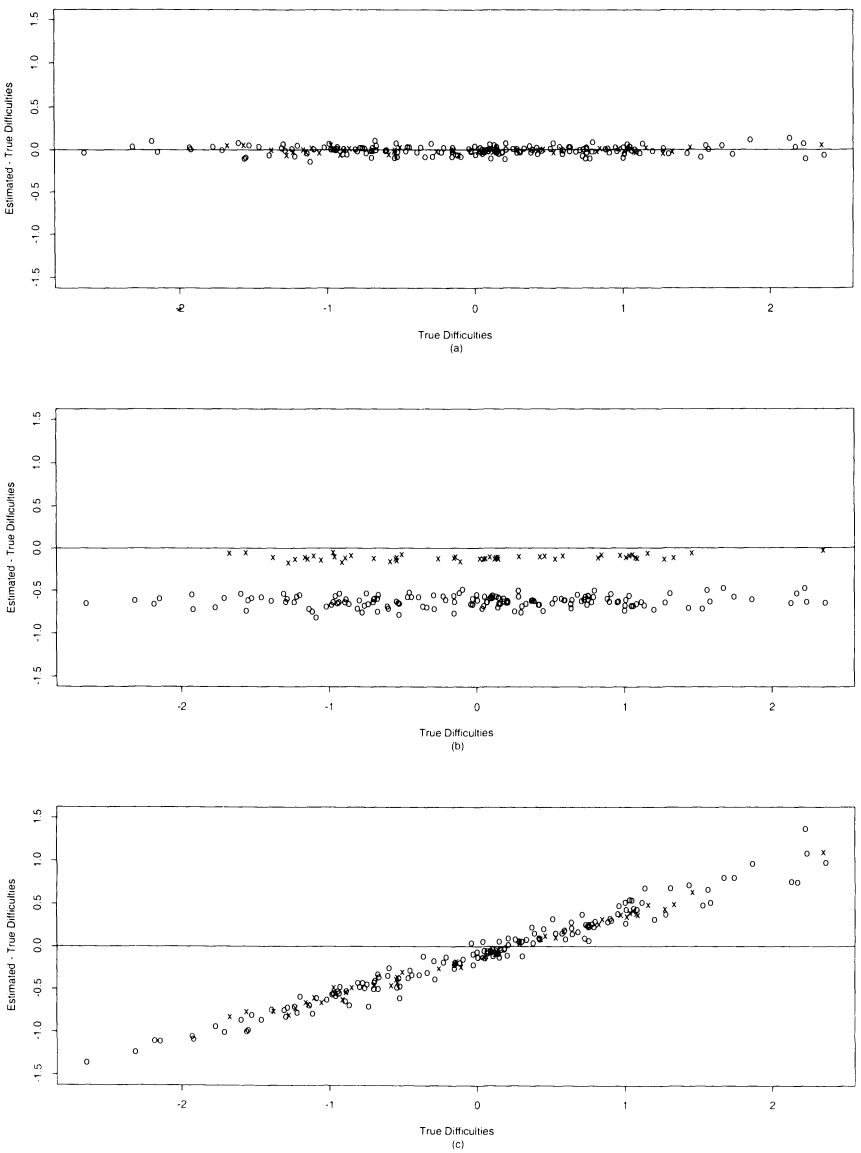

FIGURE 1. *Estimated minus true item parameter plotted against the true item parameter for three simulated data sets. In panel (a), a MAR design produces accurate inferences. In panels (b) and (c) respectively, the choice mechanism depends upon the unobserved* $\mathbf{Y}_{mis}$ *and the examinee ability and item difficulties. The solid horizontal line indicates absence of a difference. The first 50 items which are mandatory are plotted with an "x", the remaining 150 choice items with an "o".*

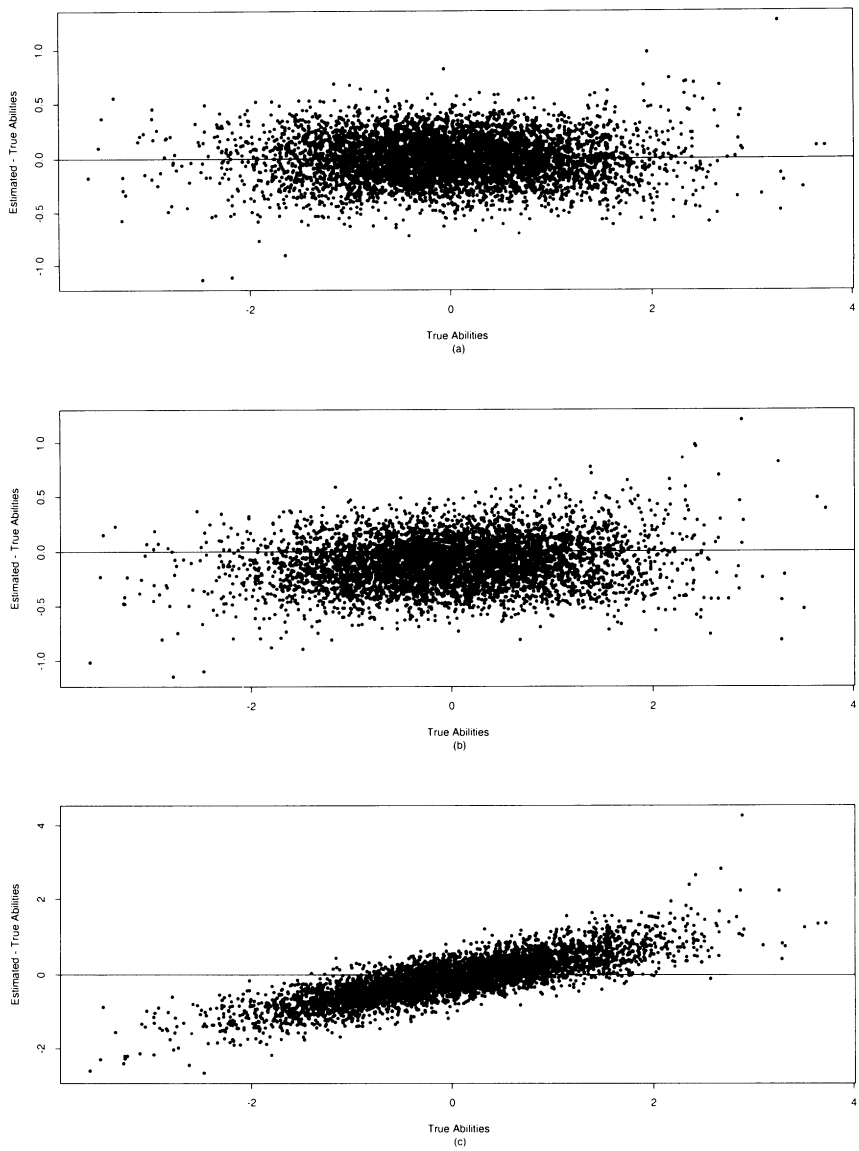

FIGURE 2. *Estimated minus true examinee ability plotted against true examinee ability for three simulated data sets. In panel (a), a MAR design produces accurate inferences for examinee abilities. In panels (b) and (c), respectively, the choice mechanism depends upon the unobserved* $\mathbf{Y}_{mis}$ *and the examinee ability and item difficulties. The solid horizontal line indicates absence of a difference. Note the difference in scale in panel (c).*

correct answer. There is an overall underestimation bias in the estimated abilities in Figure 2c, but with a trend towards decreasing bias with higher ability. This trend is complex because the high-and low-ability students had fewer choices among their 75 pairs because the items within each pair were more likely to have been both right or both wrong.

For experiment 3, Figure 1c indicates an overestimation of the difficulty parameters for high-difficulty items and a corresponding underestimation for low-difficulty ones. Because the better students tended to select the easier items, these items appear too easy, and similarly, the weaker students selecting the harder items make them appear even more difficult. The trend is similar for the choice items and the mandatory items. Abilities display a similar pattern in Figure 2c. The potential for bias in *all* of the item parameter and ability estimates makes the interpretation of IRT models dubious, even when the purpose is to study the consequences of choice (Fitzpatrick & Yen, 1995).

Conclusions

If choice is relevant, common low-dimensional IRT models may not be appropriate if, for example, students are able to distinguish favorable items, either because of specific knowledge or more general item difficulty. To be confident of standard inferences from these IRT models when applied to choice-based examinations, substantial experience under operational conditions is necessary to verify the strong assumptions in (a) and (b). Such data can be feasibly collected by spiraling different test forms which ensure that each item is mandatory for a random sample of students.

A more fundamental objection to the use of common IRT methods with choice-based examinations is the conflicting ideas motivating them. Practical IRT models typically assume that examinee performance can be characterized by a small number of abilities (often one). The choice-based examination is often motivated by the belief that examinees have numerous abilities, and the provision of choice is intended to permit each examinee to demonstrate their best abilities. If the assumptions underlying common IRT models are approximately correct, then choice-based examinations are poor designs for estimating ability. If the assumptions of the common low-dimensional IRT models are incorrect, even the strong additional assumptions (a) and (b) are insufficient to produce appropriate adjustments from IRT models. Although it is theoretically possible to develop multidimensional IRT models and corresponding models for the dependence of choice on abilities and item responses, the application of these methods in high stakes operational settings is problematic.

References

Allen, N., Holland, P., & Thayer, D. (1994). Approaches to nonignorable nonresponse with applications to selection bias. *Educational Testing Service Report*, RR–94–16.

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

Fitzpatrick, A., & Yen, W. (1995). The psychometric characteristics of choice items. *Journal of Educational Measurement, 32*, 243–259.

Little. R.J.A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley & Sons.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: LEA publishers.

Mislevy, R. J., & Sheehan, K. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661–679.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley & Sons.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*, 159–195.

Wainer, H., Wang, X., & Thissen, D. (1994). How well can we compare score on test forms that are constructed by examinees' choice? *Journal of Educational Measurement, 31*, 183–199.

Wang, X., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, 8*, 211–225.

Wang, X. (1992). *Achieving equity in self-selected subsets of test items.* Unpublished doctoral dissertation, University of Hawaii, Honolulu.

Authors

ERIC T. BRADLOW is Assistant Professor of Marketing and Statistics, Wharton School of Business, University of Pennsylvania, Suite 1400 SH-DH, 3620 Locust Walk, Philadelphia, PA 19104-6371. He specializes in Bayesian inference and statistical computing.

NEAL THOMAS is Senior Research Biostatistician, Department of Biostatistics, University of North Carolina, Chapel Hill, NC; neal thomas@ibm.net. He specializes in observational studies and missing data.

Received August 22, 1996
Revision received March 27, 1997
Accepted May 8, 1997