IS IT WRONG? HANDLING MISSING RESPONSES IN IRT

Michael J. Culbertson
Department of Educational Psychology
University of Illinois–Urbana-Champaign
Presented at the Annual Meeting of the
National Council on Measurement in Education, April 2011

Missing data present a well-known challenge to data analysis across the social sciences. Not only are many common statistical computations undefined for incomplete data, inferences based only on observed data may incur bias when the reason for missing responses is strongly associated with the values that would have been observed (Rubin, 1976), such as cases in which subjects choose not to respond when they perceive their responses to be socially undesirable. In educational assessment, response patterns may included missing data for a variety of reasons. For example, an examinee may be unwilling to guess, forget to return to a skipped item, experience fatigue, or leave the test early due to illness or some other test-unrelated reason. Even if careful test instructions encourage examinees to respond to all items, some will inevitably ignore the instructions or mistakenly leave an item blank, forcing the test analyst to determine how to handle these omitted responses.

One common method is to treat omitted responses as if they were answered incorrectly, under the logic that examinees would have indicated an answer if they had any knowledge about the item. However, by ignoring the possibility that examinees could have guessed the correct answer, this method biases ability estimates downward (Mislevy and Wu, 1996). Another common method is to treat omitted items as if they had never been administered. While this may be appropriate for blocks of omitted items at the end of a test (which examinees did not reach, presumably, due to leaving the test early or running out of time), extensive use of this strategy could encourage some examinees to artificially inflate their score by responding only to items for which they have a reasonably high degree of certainty of being correct.

In 1974, Frederick Lord proposed a modified likelihood method in which omitted items are treated as fractionally correct. In the usual likelihood function for dichotomous items under Item Response Theory (IRT):

$$L(\theta) = \prod_{i,j} P_j(\theta_i)^{u_{ij}} [1 - P_j(\theta_i)]^{1-u_{ij}} \ ,$$

where $P_j(\theta_i)$ is the probability that examinee i with ability θ_i will answer item j correctly and u_{ij} is an indicator of the examinee's actual response. Lord extended the definition of u_{ij} to be 1 if correct, 0 if incorrect, and c if examinee i omitted item j. Lord set the value of c to be the reciprocal of the number of options for

multiple-choice items. So defined, the score $\sum_j u_{ij}$ is perfectly correlated with a commonly used formula score for which examinees are penalized for incorrect answers to dissuade guessing. Lord chose to develop a method for handling omits in the formula-scoring case because it provides examinees with a rational method for choosing to omit items for which they do not know the correct answer—he explicitly avoids the sum-score case, since examinees acting in their best interest would never omit any item. However, examinees may have difficulty assessing their confidence in their knowledge about various items, particularly very young examinees. In fact, Sherriffs and Boomer (1954) found that risk-averse college students omitted items they would have answered correctly more often than non-risk-averse students, and more often than would be expected by the "perfectly rational" examinee. Alternatively, de Ayala, Plake, and Impara (2001) suggested that $c = 1/2$ be used to minimize the adverse effect of not knowing how an examinee would have answered.

Additionally, several methods have been proposed that require specification of the probability of omitting an item given item and examinee characteristics (Lord, 1983; Mislevy and Wu, 1996; Patz and Junker, 1999). Outside of IRT, analysts have achieved a fair amount of success in providing plausible values for omitted responses with the EM algorithm (Bernaards and Sijtsma, 2000) and multiple imputation (Rubin, 1987).

While the theory of the ignorability of missing data in IRT has been studied extensively (Mislevy and Wu, 1996), there have been few empirical studies of the consequences of omitted responses on ability estimates (de Ayala et al., 2001) and item parameter estimates (Holman and Glas, 2005; Finch, 2008). This simulation study examines the error incurred in ability estimates under seven treatments of omitted responses, conditional on true ability and number of items omitted. The study first examines error in ability estimates for each method under a best-case scenario (omission pattern is independent of the omitted response), and then examines error in ability estimates for the best-performing methods under a worst-case scenario (omission only of items examinees would have gotten wrong).

EMPIRICAL OMISSION PATTERNS

To gain preliminary insight into the omission behavior of examinees in a common, high-stakes educational assessment context, the omit rates for the 4th grade mathematics portion of the 2003 Massachusetts Comprehensive Assessment System (MCAS) were examined. The test included 39 items: 29 multiple-choice, 5 short answer (the examinee performs a computation and writes the result), and 5 extended response (multi-part, written response) items. Contiguous blocks of missing data at the end of the test were labeled as "not reached," and the remaining

Omits	N	
0	70,197	(94%)
1	3,066	(4%)
2	509	(0.7%)
3	191	
4	178	
5	53	
>5	225	
	74,419	

Table 1: Frequency of examinees omitting the given number of items on the MCAS.

missing data were labeled as "omitted." Examinees who omitted at least one item were labeled "omitters." While the vast majority (94%) of examinees did not omit any items, a substantial number (over 4,000) omitted at least one item (Table 1). Given the large number of examinees who do not omit any item, one could hypothesize two classes of examinees (those who omit and those who do not) and model the number of omitted items with a zero-inflated Poisson model. While most examinees who omitted responses left only a few items blank, 97 examinees omitted over a quarter of the test—raising the question of whether scores for these latter examinees should be considered at all.

For the MCAS data, three-parameter logistic (3PL) and graded response item parameters and preliminary ability estimates were obtained using marginal maximum likelihood estimation (MMLE) with expected a priori ability estimates, treating omitted and not-reached items as not administered. Item omit rates were only weakly correlated with item difficulty ($\rho = 0.19$, excluding Item 11, see Fig. 1, left). If examinees only omit items to which they do not know the answer, one would expect the average number of omitted items to decrease for increasing ability, which is observed for the omit rate among all MCAS items (Fig. 1, right). However, MCAS omit rates suggest a more complex omission mechanism when disaggregated by item type. While items for which students must actively produce a response (short answer and extended response) follow the overall trend, low-ability students are actually *less* likely to omit than high-ability students for multiple-choice items. This could result if, for example, examinees with partial knowledge (higher-ability examinees) are reluctant to guess or are more likely to skip an item intending to return to it but forget to do so, whereas low-ability examinees may simply guess when they do not know. In any case, these results suggests that simplistic accounts of omission—such as assuming examinees omit only when they have no knowledge about an item—are inadequate, and further study

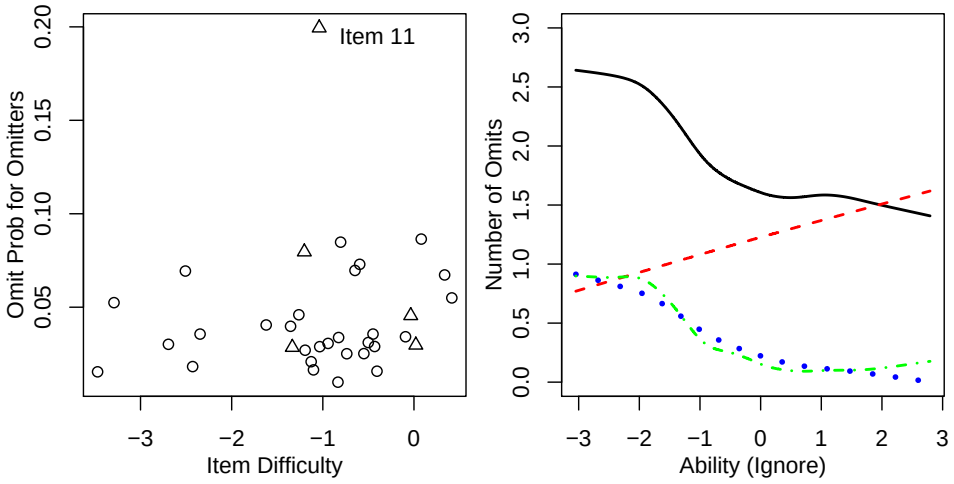

Figure 1: Left: Probability among omitters of omitting an item by item difficulty for multiple choice (circle) and short answer (triangle) items. Item 11 appears to be an outlier. Right: Average number of omitted items among omitters by estimated examinee ability for MCAS: among all items (solid), multiple choice (dashed), short answer (dotted), extended response (dash-dot). Curves obtained by smoothing splines.

of student motivation to omit is merited. Since little is know about actual omission mechanisms, this study examines the effects of omission on ability estimates conditional on ability and omit frequency.

METHODS

For the simulation study, a set of 39 item parameters were generated for the 3PL model using a standard normal distribution for the difficulty parameters, a Uniform(0.5, 2) distribution for the discrimination parameters, and a Uniform(0.1, 0.4) distribution for the pseudo-guessing parameters. Simulations were conducted in which examinees omitted 0, 1, . . ., 5 items, selected randomly with equal probability (the best-case scenario). In each simulation, 5,000 response patterns were generated for each of 41 equally-spaced points on the ability continuum from -2 to 2, inclusive. The examinee abilities were then estimated using seven treatments of omitted responses:

1. Omitted items treated as not administered
2. Omitted responses treated as incorrect
3. Omitted responses imputed by an EM algorithm

4. Abilities estimated by multiple imputation
5. Omitted items counted as fractionally correct, using 0.25 (Lord, 1974)
6. Omitted items counted as fractionally correct, using 0.5 (de Ayala et al., 2001)
7. Omitted items counted as fractionally correct, using the pseudo-guessing parameter

In the EM algorithm, the most likely response category given the item's 3PL model was imputed in the expectation step, and the examinee's ability was estimated from the full data in the maximization step. In the multiple imputation algorithm, responses were sampled from the items' 3PL models using preliminary ability estimates treating the omitted items as not administered. Abilities were estimated given the known item parameters by expected a posteriori (EAP) with a standard normal prior, maximum a posteriori (MAP), and maximum likelihood (MLE) methods. The average bias and mean squared error of estimation (MSE) were then calculated at each point on the ability continuum. Response patterns for which the MLE did not converge were excluded from the corresponding bias and MSE calculations.

RESULTS

In general, EAP and MAP estimation suffer from greater bias for very low- and high-ability examinees than MLE, but they nevertheless have smaller MSE (Fig. 2). Under EAP estimation, treating omitted responses as incorrect severely penalizes high-ability examinees (Fig. 3a). Although there is a decrease in MSE for low-ability examinee, this seeming improvement results from the cancelation of the positive bias of EAP estimation and the negative bias of the additional incorrect responses. Results for treating omitted responses as fractionally correct using 0.25 and the pseudo-guessing parameter are similar (not shown). Estimation of missing responses by the EM algorithm causes an increase in estimation error in the middle-ability range (Fig. 3b). Results for multiple imputation are similar (not shown). Treating omitted items as not administered results in only a minor increase in estimation error, concomitant with the reduced effective test length, and treating omitted items as fractionally incorrect using 0.5 results in an increase in measurement error only for high-ability examinees, though the increased error is not significantly greater than the corresponding measurement error under MLE (Fig. 4). Trends in the corresponding results for MAP are similar (not shown).

Under MLE, most treatments result in a much-smoother increase in estimation error across the ability scale with increasing number of omits than under

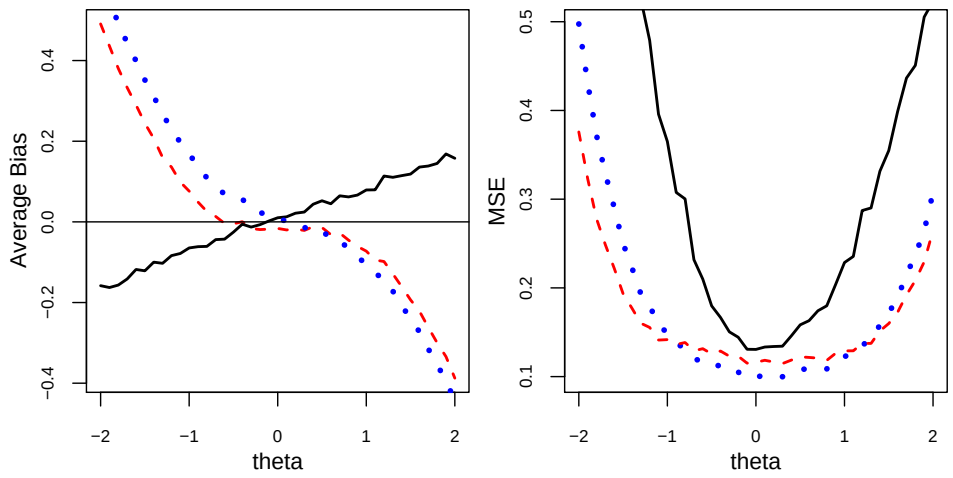

Figure 2: Average estimation bias (left) and error (right) by simulated ability for examinees with no omitted responses using MLE (solid), EAP (dashed), and MAP (dotted).

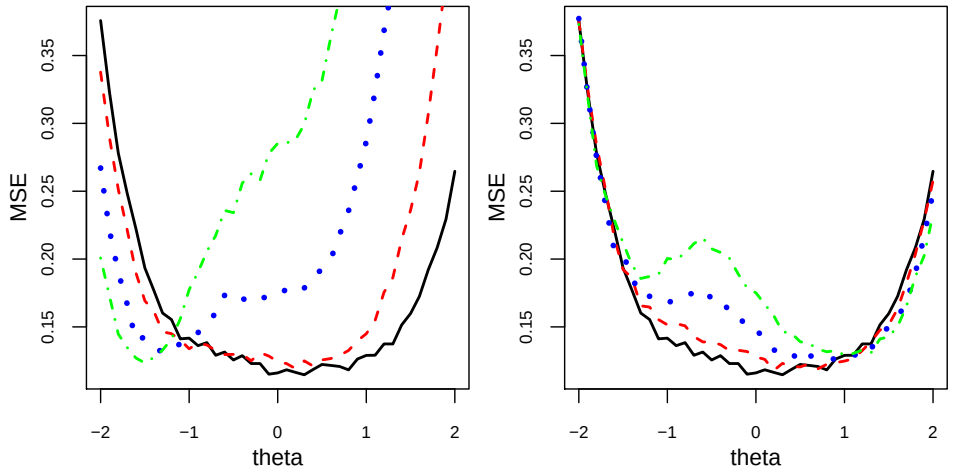

Figure 3: EAP estimation error by simulated ability when treating omitted responses as incorrect (left) and using the EM algorithm (right) for examinees with 0 (solid), 1 (dashed), 3 (dotted), and 5 (dash-dot) omitted items.

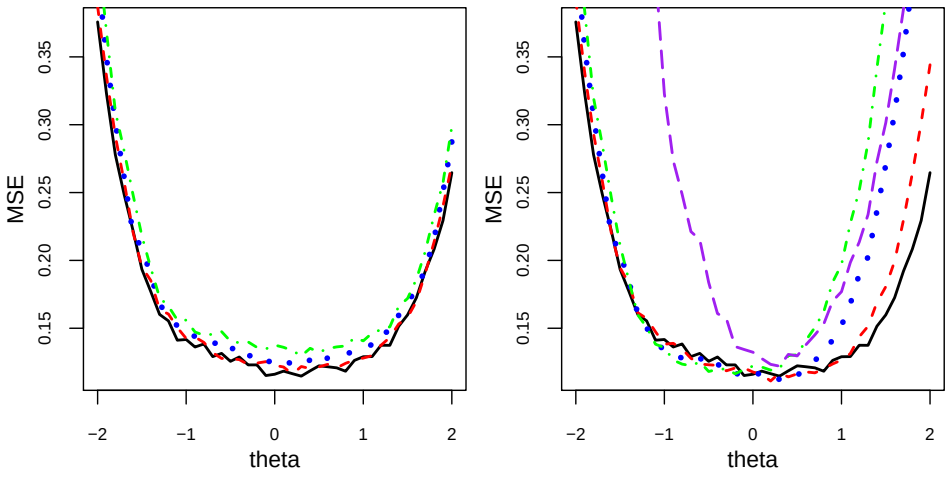

Figure 4: EAP estimation error by simulated ability when treating omitted responses as not administered (left) and fractionally correct using 0.5 (right) for examinees with 0 (solid), 1 (short dash), 3 (dotted), and 5 (dash-dot) omitted items. MLE estimation error for 5 omitted items treated as fractionally correct is provided for comparison (long dash, right).

EAP (Fig. 5). Similar to the EAP results, treating omitted items as not-administered results in only a very slight increase in error, and the error when treating omitted items as fractionally correct using 0.5 is nearly identical to the error with no omitted items.

The preceding simulations assume a best-case scenario in which the omission mechanism is independent of the omitted response. Under some conditions, such as treating omitted responses as not-administered, devious examinees could artificially inflate their score by responding only to items for which they are reasonably confident of having the correct response. To judge the effect of such behavior, follow-up simulations in a worst-case scenario were conducted in which omitted responses were sampled only from those items an examinee answered incorrectly, assuming that examinees have perfect assessment of their knowledge about each item. (If a high-ability examinee had fewer incorrect answers than the specified number of omitted responses, the omits remaining after selecting all incorrect items were sampled randomly from among the correct items.) The follow-up simulations were conducted only treating omitted items as not-administered and fractionally correct using 0.5, under EAP and MLE. As would be expected, both treatments result in a significant positive bias for "devious" examinees, particularly at the upper end of ability (Fig. 6).

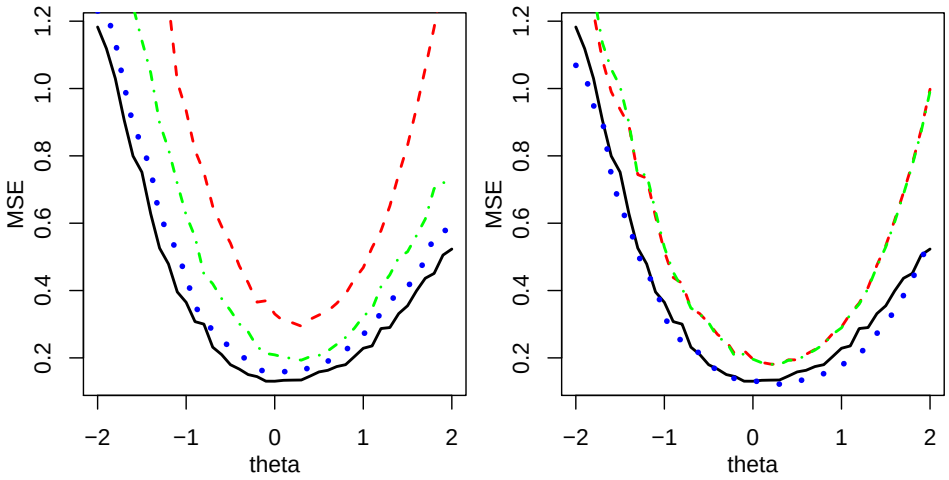

Figure 5: MLE estimation error by simulated ability for examinees with 5 omitted items. Solid line indicates MLE estimation error for examinees with no omitted items. Left: omits treated as incorrect (dashed), omits treated as not administered (dotted), multiple imputation (dash-dot); right: omits treated as fractionally correct using 0.25 (dashed), 0.5 (dotted), pseudo-guessing parameter (dash-dot).

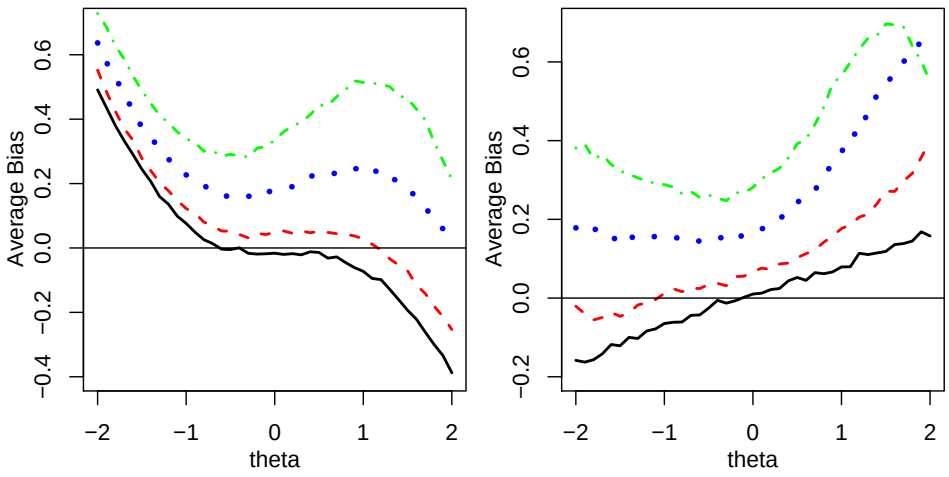

Figure 6: Average bias when examinees omit only items they would have gotten incorrect for examinees with 0 (solid), 1 (short dash), 3 (dotted), and 5 (dash-dot) omitted items, treating omits as not administered with EAP (left) and as fractionally correct using 0.5 with MLE (right).

DISCUSSION

In the current environment of high-stakes educational assessment, accurate measurement of student ability is crucial for fair implementation of educational accountability programs. However, the most common treatment of omitted responses, coding them as incorrect, negatively biases ability estimates. Even when examinees are encouraged to respond to all items, a significant minority still omit some items, and not enough is known about the mechanism that produces these omissions. Much more research is necessary in the psychology of item omission, including further studies in the line of Sherriffs et al. (1954) and those proposed by Mislevy and Wu (1996).

When omission is independent of the omitted responses, treating omitted responses as not-administered or half-correct has the least impact on measurement error. It is particularly notable that the increased error for these two treatments is significantly lower than the increase due to negative bias from treating omitted responses as incorrect. Unfortunately, due to external pressures to produce strong test results, these treatments are likely not feasible in large-scale educational measurement due to their susceptibility to abuse. More needs to be known about the mechanism producing omissions in order to devise better statistical strategies to account for missing responses in educational assessment data. This could be investigated with studies similar to Sherriffs and Bloomer (1954) and the study proposed by Mislevy and Wu (1996) in which students are given a test in which they may omit items and are then asked to state the answers they would have given to omitted items. As with Sherifs and Bloomer (1954), the association of omission rates and various psychological properties (such as risk aversion) could be investigated in detail, along with how students respond to various test instruction conditions. In the end, until more is known about how examinees choose to omit items, the most pragmatic choice in high-stakes testing may be to continue to treat omitted responses as incorrect and to encourage examinees to respond to all items.

REFERENCES

Bernaards, Coen A. and Sijtsma, Klaas (2000). "Influence of Imputation and EM Methods on Factor Analysis when Item Nonresponse in Questionnaire Data is Nonignorable." *Multivariate Behavioral Research,* **35**, 321–364.

de Ayala, R. J.; Plake, Barbara S.; and Impara, James C. (2001). "The Impact of Omitted Responses on the Accuracy of Ability Estimation in Item Response Theory." *Journal of Educational Measurement,* **38**, 213–234.

Finch, Holmes (2008). "Estimation of Item Response Theory Parameters in the Presence of Missing Data." *Journal of Educational Measurement,* **45**, 225–245.

Holman, Rebecca and Glass, Cees A. W. (2005). "Modelling Non-Ignorable Missing-Data Mechanisms with Item Response Theory Models." *British Journal of Mathematical and Statistical Psychology,* **58**, 1–17.

Lord, Frederic M. (1974). "Estimation of Latent Ability and Item Parameters When There are Omitted Responses." *Psychometrika,* **39**, 247–264.

Mislevy, Robert J. and Wu, Pao-Kuei (1996). *Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing*, Educational Testing Service report no. RR-96-30-ONR.

Rubin, Donald B. (1976). "Inference and Missing Data." *Biometrika*, **63**, 581–592.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Sherriffs, Alex C. and Boomer, Donald S. (1954). "Who Is Penalized by the Penalty For Guessing." *Journal of Educational Psychology,* **45**, 81–90.