Journal of Educational Measurement
Fall 2001, Vol. 38, No. 3, pp. 213–234

The Impact of Omitted Responses on the Accuracy of Ability Estimation in Item Response Theory

R. J. De Ayala
Barbara S. Plake
James C. Impara
University of Nebraska-Lincoln

Practitioners typically face situations in which examinees have not responded to all test items. This study investigated the effect on an examinee's ability estimate when an examinee is presented an item, has ample time to answer, but decides not to respond to the item. Three approaches to ability estimation (biweight estimation, expected a posteriori, and maximum likelihood estimation) were examined. A Monte Carlo study was performed and the effect of different levels of omissions on the simulee's ability estimates was determined. Results showed that the worst estimation occurred when omits were treated as incorrect. In contrast, substitution of 0.5 for omitted responses resulted in ability estimates that were almost as accurate as those using complete data. Implications for practitioners are discussed.

For a variety of reasons, an examinee's response vector may not contain responses to each item. For example, the items not presented in an adaptive test or the non-common items in a common-item equating design will have responses for only a subset of examinees. In both of these examples, the test administration involves a decision not to present certain items to all examinees. Using Little and Rubin's (1987) terminology, these nonresponses represent conditions in which the missingness process may be ignored for purposes of ability estimation (Mislevy & Wu, 1988; Mislevy & Wu, 1996). In contrast, nonresponses for "not-reached" item(s) occur because an examinee has insufficient time to even consider answering the item(s). These not-reached items can be identified as collectively occurring at the end of an examination, assuming the examinee responds to the test items in serial order. Lord (1980) stated that in practice these not-reached items may be ignored for ability estimation because they contain no readily quantifiable information about the examinee's ability. Augmenting this perspective, Mislevy and Wu (1996) outlined the conditions in which not-reached items may represent ignorable missing data. Another source of missing data occurs because examinees have the capability of choosing not to respond to certain questions on an examination. Such (intentionally) omitted responses represent nonignorable missing data (Lord, 1980; Mislevy & Wu, 1988; Mislevy & Wu, 1996). This study investigated the effect on an examinee's ability when an examinee is presented an item, has ample time to answer the item, but decides not to respond to the item.

It is reasonable to believe that, in general, an examinee who omits responding to an item does so because the examinee believes that he or she does not know the answer. A highly proficient individual, by virtue of his or her ability, may be more

likely to realize that he or she does not know the answer to an item better than a less proficient examinee. Therefore, the highly proficient examinee may have a greater tendency to omit items for which he or she does not know the answers than does a less proficient examinee. In addition, the highly proficient examinee may tend to omit responses at a lower rate than does a less proficient examinee. Conversely, a less proficient examinee may not be able to make the distinction that he or she knows the answer to a question as well as does a highly proficient examinee and as a consequence may omit items that may have been answered correctly if he or she had decided to respond (cf. Stocking, Eignor, & Cook, 1988; Wainer & Thissen, 1994). Clearly, in the context of ability estimation, omitted responses are not ignorable because the act of omission is related, in part, to the examinee's ability. Lord (1980) has argued that omitted responses may not be ignored because an examinee who understands ability estimation in the context of item response (IRT) could obtain as high an ability estimate as he or she wished simply by answering only those items he or she has confidence in answering correctly. This idea has found some support in Wang, Wainer, and Thissen's (1995) study on examinee item choice.

A number of different ability estimation approaches in IRT have different advantages and disadvantages. It might be expected that the effect of omitted responses on ability estimation may vary as a function of estimation approach. For instance, if a Bayesian-based method is used (e.g., expected a posteriori), then the regression toward the mean phenomenon inherent in a Bayesian approach might be expected to compensate to some extent for the potential underestimation of less proficient examinees and the overestimation of highly proficient examinees who do not respond to all the items. In contrast, a maximum likelihood-based approach might be expected to show the aforementioned biases. A procedure proposed by Mislevy and Bock (1982), biweight ability estimation, was developed to provide robust ability estimation using maximum likelihood. With this method, the likelihood is modified to weight items closer to the examinee's proficiency more than those further away ("trimming"). Weighting the items appropriately may provide a means of compensating for the expected biases and result in a more accurate ability estimate than would be obtained using a nonweighted maximum likelihood approach. An alternative approach to dealing with missing data is based on Lord (1974). This method involves the assignment of a fractionally correct value equal to the reciprocal of the number item alternatives (i.e., the random guessing value) to the omitted item(s). This latter method assumes that examinees omit items if their chances of correctly responding would have been equal to random guessing. In addition, this approach assumes that both highly and less proficient examinees can be treated the same. However, this assumption may not be tenable; Stocking, Eignor, and Cook (1988) have shown that the rates of omission vary as a function of ability. Moreover, Mislevy and Wu (1988) have stated that the tendency to omit can be associated with personality characteristics and demographic variables, as well as ability level. Therefore, differential omission rates may not be compensated for using Lord's approach for proficiency estimation. This study investigated the effect of omitted responses on three approaches to ability estimation: biweight estimation, expected a posteriori (EAP), and maximum likelihood estimation

(MLE). For the biweight method, different levels of trimming were examined to determine if the effect of omitted responses on ability estimates could be ameliorated. For the Bayesian method, EAP, we examined whether the use of prior information would compensate for the effect omitted responses have on ability estimation. For the MLE method, Lord's approach of replacing the omitted responses with the reciprocal of the number of alternatives was studied. Moreover, for all three methods, omitted responses were also treated as incorrect or ignored and the effect on ability estimation examined.

Method

Ability Estimation Methods

Ability estimation typically has used either MLE or a Bayesian approach such as maximum a posteriori (MAP or Bayes Modal Estimate) or EAP (Bayes Mean Estimate). The former two algorithms are iterative techniques, while EAP (Bock & Mislevy, 1982) is noniterative and is based on numerical quadrature methods. Unlike MLE ability estimates, EAP ability estimates may be obtained for all response patterns, including zero and perfect score patterns. Although MAP proficiency estimates also exist for all response patterns, they suffer from greater regression towards the mean than do EAP estimates (Bock & Mislevy, 1982; Mislevy & Bock, 1990). The EAP estimate ($\hat{\theta}$) of an examinee's proficiency, θ, after n items have been administered is given by

$$\hat{\theta}_n = \frac{\sum_{k=1}^{q} X_k L_n(X_k) A(X_k)}{\sum_{k=1}^{q} L_n(X_k) A(X_k)}, \tag{1}$$

and its posterior standard deviation is

$$\mathrm{PSD}(\hat{\theta}) = \sqrt{\frac{\sum_{k=1}^{q} (X_k - \hat{\theta}_n)^2 L_n(X_k) A(X_k)}{\sum_{k=1}^{q} L_n(X_k) A(X_k)}}, \tag{2}$$

where X_k is one of q quadrature points, $A(X_k)$ is the corresponding quadrature weight, and $L_n(X_k)$ is the likelihood function of X_k given the response pattern $\{x_1, x_2, \ldots, x_n\}$. For example, if the probability of a correct response by an individual with proficiency θ to a dichotomously scored item i with location b_i, discrimination a_i, and pseudo-guessing parameter c_i is given by the three-parameter logistic (3-PL) model

$$p(x_i = 1 \mid \theta) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta - b_i)}}, \tag{3}$$

then the likelihood of θ given the response pattern $\{x_1, x_2, \ldots, x_n\}$ is

$$L_n(\theta) = \prod_{i=1}^{n} p(x_i = 1 \mid \theta)^{x_i} [1 - p(x_i = 1 \mid \theta)]^{(1-x_i)}. \tag{4}$$

MLE uses a gradient approach for determining the location of the maximum of Equation 4 (i.e., the value that maximizes the likelihood). This location is taken as the examinee's $\hat{\theta}$. In practice, the natural log of Equation 4 typically is used. Given some estimate of an examinee's θ, $\hat{\theta}^t$, the estimate is refined by examining the average rate of change of the function with respect to a particular point. Technically, this refinement takes the form of a ratio (Δ) of the first derivative to the second derivative of the log likelihood function (Lord, 1980):

$$\frac{\partial \ln L}{\partial \theta} = \frac{\sum_{i=1}^{n} a_i (x_i - p_i)(p_i - c_i)}{p_i (1 - c_i)} \tag{5}$$

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{\sum_{i=1}^{n} a_i^2 (x_i c_i - p_i^2)(p_i - c_i)(1 - p_i)}{p_i^2 (1 - c_i)^2}, \tag{6}$$

where p_i is defined by Equation 3 given the appropriate item parameters and current θ. Therefore, the refinement of θ at the $t + 1$ iteration is given by

$$\hat{\theta}^{t+1} = \hat{\theta}^t - \Delta = \hat{\theta}^t - \frac{\dfrac{\partial \ln L}{\partial \theta}}{\dfrac{\partial^2 \ln L}{\partial \theta^2}}. \tag{7}$$

Iterations continue until $\hat{\theta}^{t+1}$ is considered to be equivalent to $\hat{\theta}^t$ to some degree of accuracy. At this point, the examinee's $\hat{\theta}$ is taken to be $\hat{\theta}^{t+1}$. The estimate of the standard error of estimate for $\hat{\theta}$ is

$$\mathrm{SEE}(\hat{\theta}) = \sqrt{\frac{1}{\sum_{i=1}^{n} \dfrac{[a_i (1 - p_i)(p_i - c_i) / (1 - c_i)]^2}{p_i (1 - p_i)}}}, \tag{8}$$

and p_i is from the final iteration.

Mislevy and Bock (1982) introduced a modification of MLE to reduce its sensitivity to responses that are inconsistent with an IRT model (e.g., Equation 3). An example of such a "response disturbance" would be an incorrect response to an "easy" item by a high-ability examinee. Their modification involved the application of Tukey's biweight to the estimation of θ. The biweight is primarily inversely related to the distance between an item's location and the examinee's biweight $\hat{\theta}$. The closer the item is to the examinee's $\hat{\theta}$, the greater the weight given to the item; the further away the item is from the examinee's $\hat{\theta}$, the less the weight that is given to the item (the weight is defined on the following page). In short, unexpected

responses are given less weight than responses that are consistent with the model. Estimation proceeds as in Equation 7 except that the ratio of derivatives is modified to include "item weights" (W_i) that are iteration specific and reflect the amount of trimming, C:

$$\Delta = \frac{\sum W_i^t a_i (x_i - p_i^t)}{-\sum W_i^t a_i^2 p_i^t (1 - p_i^t)}, \tag{9}$$

where $W_i^t = \{1 - [a_i(b_i - \hat{\theta}^t)/C]^2\}^2$ if $\mid a_i(b_i - \hat{\theta}^t)/C \mid \leq 1$, otherwise $W_i^t = 0$, and p_i is defined by Equation 3 given the appropriate item parameters and current $\hat{\theta}$. If $(a_i(b_i - \hat{\theta}^t)/C) > 1$, then the item weight is zero and the item is effectively "removed" or "trimmed." C is an arbitrary constant that specifies, as a function of the logit, the amount of trimming to be done. Therefore, there is an inverse relationship between C and the amount of trimming. According to Mislevy and Bock (1982), smaller values of C implement heavy trimming (e.g., $C = 2$), whereas larger values such as $C = 6$, reflect lighter trimming. As was the case with MLE, iterations continue until $\hat{\theta}^{t+1}$ is considered to be equivalent to $\hat{\theta}^t$ to some degree of accuracy. This $\hat{\theta}^{t+1}$ is taken as the examinee's $\hat{\theta}$. The estimate of the standard error of estimate for $\hat{\theta}$ is appropriately modified to represent the level of trimming performed:

$$\text{SEE}(\hat{\theta}) = \sqrt{-\sum W_i a_i^2 p_i (1 - p_i)}, \tag{10}$$

where W_i and p_i are from the final iteration.

Data Generation

The simulation data were modeled on an empirical data set that consisted of 24,546 examinees and had been calibrated using the 3-PL model. Of these examinees, 6,515 had response vectors that contained a combination of correct, incorrect, and omitted responses to 39 items. The average number of correct responses for these 6,515 examinees was 19.184 (median = 18) with a standard deviation of 8.089 (minimum score = 1, maximum score = 38, skew = 0.300). For these latter examinees, the average number of items omitted was 2.224 (median = 1) with a standard deviation of 2.851 (minimum number omitted = 1, maximum number omitted = 35, skew = 4.251). Among the 6,515 examinees, 96% omitted eight items or less. Because an examinee may omit an item for many different reasons (e.g., knowledge of the answer, self-confidence, risk-aversion, test-wiseness, metacognitive factors) and there were no explicit measures of these factors, it was decided not to use a parametric approach for modeling the empirical data. Because the omission pattern across ability differed for persons who responded correctly versus incorrectly to an item, a pair of contingency tables was created for each item using 6,515 examinees who had response vectors containing correct, incorrect, and omitted responses. Each contingency table consisted of a two-level response-type variable versus an ability-measure variable. For one table, the response-type variable consisted of response omission or responding incorrectly to the item; whereas for the other table, for that item the response-type variable consisted of omitting a response

or correctly responding to the item. The ability-measure variable consisted of 10 four-item fractiles of the number correct score (0–3, 4–7, etc.). By using 10 four-item fractiles in lieu of deciles, it was felt that we would avoid having some fractiles that consisted of a relatively large range of number-correct scores and others that consisted of one or two number-correct scores. Based on these tables, the proportion of individuals omitting a response to an item conditional on the fractile was calculated.

The simulated data were generated on the basis of Equation 3, and the item-parameter estimates of the empirical data were treated as known. For each 0.1 of logit from −2.0 to 2.0 (inclusive) 1,000 θs were generated for a total of 41,000 simulees. For each simulee, the probability of a correct response was calculated according to Equation 3 and compared to a uniform random number [0,1]. If the random number was less than or equal to the probability of a correct response, the response was coded as "1" for correct, "0" otherwise. To generate the omission data, the number-correct score for each simulee was determined and the simulee assigned to 1 of the 10 fractiles. For each item the correctness of the simulee's response was used to determine which of the two contingency tables for the item should be used. Based on the simulee's fractile assignment, the appropriate relative frequency of omission was compared to a uniform random number [0,1]. If the uniform random number was less than or equal to the relative frequency for omission, conditional on the simulee's fractile, then the response was changed to be an omission; otherwise the simulee's response was not changed. For example, for an item the relative frequency of omission for an examinee in the third fractile might be 0.42 if the simulee responded incorrectly to the item and 0.11 if the simulee responded correctly. If the simulee's generated response to the item was incorrect, then a uniform random number would be generated and compared to 0.42. If this random number was 0.3, for instance, then the simulee's incorrect response to this item would be changed to reflect that it had been omitted. This process was repeated for each of the 39 items and for all simulees. Therefore, each simulee had a response vector of correct and incorrect responses (i.e., the complete vector) and a response vector of correct, incorrect, and omitted responses (i.e., the omission vector).

The study was conducted in two phases. Phase 1 was an exploratory study comparing the various estimation methods under different conditions. Phase 2 was based on Phase 1 results and examined a modified EAP approach.

Phase 1

Factors

The Biweight, EAP, and MLE estimation methods were investigated. For the Biweight method, five different levels of trimming were examined ($C = 2, 4, 6, 8,$ and 10); for EAP, two different levels of quadrature points (10 and 20 points); and for MLE, the omitted responses were replaced with the reciprocals 4 and 7 (7 was approximately equal to the reciprocal of the median c value; this factor was called Nalt for number of alternatives). The C values of 2 through 10 reflect a range of trimming from severe to light, respectively, and were selected to parallel those exemplified by Mislevy and Bock (1982). In addition, for the three methods,

omitted responses were treated as Incorrect as well as Ignored for ability estimation. The ignored condition was included to determine the impact of ignoring nonignorable missing data. Each simulee's ability was estimated using each method. For each method, each simulee had two $\hat{\theta}$s: one based on the simulee's complete vector ($\hat{\theta}_c$) and the other $\hat{\theta}$ based on the simulee's omission vector ($\hat{\theta}_o$). All methods used Equation 3.

Each level of the ability estimation methods was crossed by the number of items omitted in the response vector (Nomitted). Nomitted consisted of four levels: two, four, six, and eight omitted responses (for the simulated data, the cumulative percent for omitting eight items or less was 99.5%). These four levels of Nomitted, two, four, six, and eight, represent 5.1%, 10.3%, 15.4%, and 20.5% of the test length, respectively.

Analysis

Descriptive statistics were calculated on the item parameters and ability estimates. Fidelity coefficients were obtained. Each proficiency estimate's Root Mean Square Error (*RMSE*) and Bias were calculated:

$$RMSE(\theta_k) = \sqrt{\frac{\sum (\hat{\theta} - \theta_k)^2}{n_k}}; \tag{11}$$

$$\text{Bias}(\theta_k) = \frac{\sum (\hat{\theta} - \theta_k)}{n_k}, \tag{12}$$

where $\hat{\theta}$ is the proficiency estimate based on one of the estimation methods using either the complete or omission vectors, θ_k is the simulee's proficiency at logit k ($-2.0, -1.9, -1.8, \ldots, 2.0$), and n is the number of simulees at logit k.

RMSE and Bias were calculated separately for the complete vectors and omission vectors. Because *RMSE*s for the complete vectors represented how well the simulees could be estimated on the basis of complete response data, the *RMSE*s for the omission vectors were compared to the corresponding *RMSE*s for the complete vectors; this was also true for Bias. These differences between the *RMSE* for the omission and complete vectors as well as for Bias were examined graphically for each condition. All statistics were calculated using convergent cases.

Programs

Biweight, EAP, and MLE programs were written to perform the ability estimation. A program to calculate *RMSE* and Bias also was written.

Phase 2

Based on the results of Phase 1, another condition was implemented. The same analysis measures and data used in Phase 1 were used for Phase 2. The pattern of results of MLE using 4 and 7 as the number of alternatives suggested that using a value less than 4 as the number of alternatives may be productive. A value of 2 was chosen because 2 bisects the probability space so that at worst the probability of a correct response used in the construction of the likelihood is off by 0.5, irrespective

TABLE 1
Item Pool Descriptive Statistics

Statistics	Item parameter			Item parameter intercorrelation			
	a	*b*	*c*		*a*	*b*	*c*
M	0.8866	−0.5547	0.1644	*a*			
Median	0.8815	−0.5350	0.1347	*b*	0.3505		
Standard deviation	0.2621	0.8194	0.1271	*c*	0.3422	0.5918	
Minimum	0.4168	−2.2552	0.0000				
Maximum	1.5572	1.2916	0.4388				

of what the estimated probability would have been if the examinee had chosen to answer the omitted item. In contrast, if the number of alternatives is specified as 4, then the probability of a correct response used in estimation would be off by 0.75 if the examinee's probability of a correct response would have been 1.0 if he or she had answered the omitted item. Conversely, the probability of a correct response used in estimation would be, at best, off by 0.25 if the examinee's probability of a correct response would have been 0 for the omitted item. To the extent that the random-guessing model is incorrect for an examinee, then the magnitude of these errors will tend to be larger when the number of alternatives is specified to be larger than 2. Metaphorically, the use of 2 is analogous to dividing one's money equally across many numbers at the roulette table versus placing all of one's money on a single number.

EAP was selected as the ability estimation method for Phase 2 because it is a noniterative method for which finite ability estimates are always available and because, on average, its performance was better than MLE. Because the comparison of EAP results using 10 quadrature points were very similar to those using 20 points, and both MULTILOG (Thissen, 1991) and BILOG (Mislevy & Bock, 1990) use 10 quadrature points as default for EAP estimation, the number of quadrature points used in Phase 2 was 10. For comparison purposes, Nalt values of 4 and 7 also were used.

Results

Item pool

Table 1 contains descriptive statistics for the item pool used as part of the data generation. The item locations were distributed between −2.26 and 1.3 and centered at −0.5547, with an average item discrimination of 0.8866. The correlations between the number of times an item was omitted and item discrimination, location, and intercept were 0.0338, −0.3291, and 0.3509, respectively. The maximum test information was approximately 5.19 and was located at −0.2185.

The four levels of Nomitted consisted of 9,713 simulees that omitted two items, 6,948 that omitted four items, 2,229 that omitted six items, and 431 that omitted eight items. For these levels, the average trait values were $\bar{\theta}_2 = 0.3604$ (SD = 1.1332), $\bar{\theta}_4 = -0.3335$ (SD = 1.0830), $\bar{\theta}_6 = -0.8139$ (SD = 0.9029), and $\bar{\theta}_8 = -1.0694$ (SD = 0.7721).

TABLE 2
Descriptive Statistics and Fidelity Coefficients—Biweight

Omit	Level	Nomitted	$r\theta\hat{\theta}_c$	$r\theta\hat{\theta}_o$	$r\hat{\theta}_c\hat{\theta}_o$	$\bar{\hat{\theta}}_c$	$\bar{\hat{\theta}}_o$	Nonconvergent case: Complete	Nonconvergent case: Omit
Ignored	$C=2$	2	0.9185	0.9066	0.9774	0.3822	0.2984	0	0
		4	0.9099	0.8931	0.9768	−0.2363	−0.1888	0	0
		6	0.8690	0.8457	0.9671	−0.6432	−0.5305	0	0
		8	0.8319	0.8178	0.9466	−0.9106	−0.7606	0	0
	$C=4$	2	a	0.9148	0.9914	a	0.3991	0	0
		4	a	0.8997	0.9871	a	−0.1570	0	0
		6	a	0.8527	0.9769	a	−0.5445	0	0
		8	a	0.8248	0.9598	a	−0.8009	0	0
	$C=6$	2	a	0.9144	0.9928	a	0.4297	0	0
		4	a	0.8998	0.9881	a	−0.1466	0	0
		6	a	0.8530	0.9776	a	−0.5470	0	0
		8	a	0.8247	0.9615	a	−0.8115	0	0
	$C=8$	2	a	0.9139	0.9930	a	0.4417	0	0
		4	a	0.8996	0.9882	a	−0.1425	0	0
		6	a	0.8529	0.9777	a	−0.5479	0	0
		8	a	0.8244	0.9619	a	−0.8155	0	0
	$C=10$	2	a	0.9136	0.9931	a	0.4475	0	0
		4	a	0.8994	0.9883	a	−0.1405	0	0
		6	a	0.8528	0.9778	a	−0.5483	0	0
		8	a	0.8243	0.9620	a	−0.8175	0	0
Incorrect	$C=2$	2	0.9185	0.9066	0.9772	0.3822	0.1993	0	0
		4	0.9099	0.8945	0.9760	−0.2363	−0.4173	0	0
		6	0.8690	0.8478	0.9597	−0.6432	−0.8851	0	0
		8	0.8319	0.8092	0.9297	−0.9106	−1.2284	0	0
	$C=4$	2	a	0.9146	0.9919	a	0.2333	0	0
		4	a	0.9000	0.9873	a	−0.4634	0	0
		6	a	0.8518	0.9755	a	−0.9634	0	0
		8	a	0.8218	0.9578	a	−1.3250	0	0
	$C=6$	2	a	0.9146	0.9930	a	0.2319	0	0
		4	a	0.8995	0.9879	a	−0.4791	0	0
		6	a	0.8496	0.9766	a	−0.9823	0	0
		8	a	0.8200	0.9604	a	−1.3488	0	0
	$C=8$	2	a	0.9143	0.9930	a	0.2311	0	0
		4	a	0.8990	0.9878	a	−0.4848	0	0
		6	a	0.8485	0.9767	a	−0.9893	0	0
		8	a	0.8189	0.9608	a	−1.3577	0	0
	$C=10$	2	a	0.9141	0.9929	a	0.2307	0	0
		4	a	0.8988	0.9877	a	−0.4875	0	0
		6	a	0.8480	0.9766	a	−0.9925	0	0
		8	a	0.8182	0.9610	a	−1.3620	0	0

a = Because estimation converged for all cases at this level and no trimming was done on the complete vector, the $r\theta\hat{\theta}_c$s and the $\bar{\hat{\theta}}_c$s are the same for a given level of Nomitted across C levels as well as ignoring omits or treating them as incorrect.

TABLE 3
Descriptive Statistics and Fidelity Coefficients—EAP

Omit	Level	Nomitted	$r\theta\hat{\theta}_c$	$r\theta\hat{\theta}_o$	$r\hat{\theta}_c\hat{\theta}_o$	$\hat{\theta}_c$	$\hat{\theta}_o$
Ignored	10 Points	2	0.9183	0.9127	0.9929	0.3934	0.4721
		4	0.9097	0.8991	0.9880	−0.2457	−0.1433
		6	0.8682	0.8524	0.9773	−0.6664	−0.5711
		8	0.8328	0.8249	0.9624	−0.9428	−0.8538
	20 Points	2	0.9184	0.9129	0.9931	0.3935	0.4720
		4	0.9099	0.8992	0.9883	−0.2451	−0.1430
		6	0.8692	0.8531	0.9779	−0.6656	−0.5701
		8	0.8325	0.8245	0.9626	−0.9418	−0.8527
Incorrect	10 Points	2	a	0.9133	0.9923	a	0.2358
		4	a	0.8982	0.9867	a	−0.5097
		6	a	0.8477	0.9749	a	−1.0314
		8	a	0.8165	0.9602	a	−1.4144

a = $r\theta\hat{\theta}_c$s and the $\bar{\theta}_c$s are the same for a given level of Nomitted across Quadrature Point levels and treating omits as incorrect or ignoring them.

Phase 1

Table 2 shows the fidelity coefficients as well as the intercorrelation between the Biweight ability estimates based on the complete vectors and the omission vectors; $\bar{\hat{\theta}}_c$ and $\bar{\hat{\theta}}_o$ represent the mean estimates using the complete and omission vectors, respectively. For all levels of the trimming factors as well as treating omits as incorrect or ignoring them, the $r_{\theta\hat{\theta}_c}$s were greater than the $r_{\theta\hat{\theta}_o}$s for corresponding Nomitted levels. As would be expected, for a given trim level, an increase in the number of omits resulted in a decrease in the fidelity coefficients. These decreases were similar across the trim-factor levels with a difference of approximately 0.09 between the largest and smallest fidelity coefficients. Although the $r_{\theta\hat{\theta}_o}$s were greater for the lower trim levels (i.e., $C = 4, 6, 8, 10$) than for the highest trim level (i.e., $C = 2$) for a given level of omission, the differences between corresponding $r\theta\hat{\theta}_o$s for similar Nomitted levels were slight. Reducing the amount of trimming from $C = 4$ to $C = 10$ tended to produce $r_{\theta\hat{\theta}_o}$s that were similar in magnitude for a given level of omission. Comparing the $r_{\hat{\theta}_c\hat{\theta}_o}$s across trim levels for corresponding levels of Nomitted showed that the $r_{\hat{\theta}_c\hat{\theta}_o}$s tended to increase as less trimming was used on the $\hat{\theta}_c$s. In general, for a given trim level, the $r_{\hat{\theta}_c\hat{\theta}_o}$s decreased with increasing level of Nomitted, although they were still greater than 0.94. Comparing Omits Treated as Incorrect responses with the Ignored Omits level appeared to have little impact on the magnitude of the correlation coefficients, although the $\hat{\theta}_o$ were consistently less when omits were treated as incorrect than when they were ignored.

The EAP results are displayed in Table 3. Comparison of the $r_{\theta\hat{\theta}_o}$s across the number of quadrature point levels showed that for corresponding Nomitted levels these correlations varied by less than 0.0007. The fidelity coefficients involving the complete vectors ($r_{\theta\hat{\theta}_c}$) for 10 quadrature points differed from the $r_{\theta\hat{\theta}_c}$s for 20

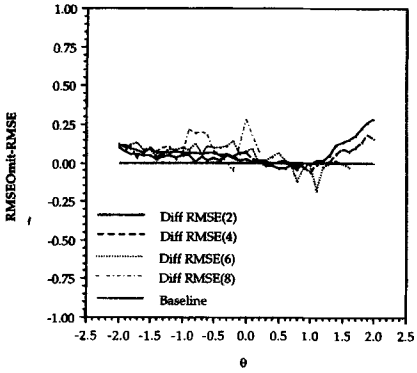
TABLE 4
Descriptive Statistics and Fidelity Coefficients—MLE

Level	Nomitted	$r\theta\hat{\theta}_c$	$r\theta\hat{\theta}_o$	$r\hat{\theta}_c\hat{\theta}_o$	$\bar{\hat{\theta}}_c$	$\bar{\hat{\theta}}_c$	Nonconvergent case Complete	Nonconvergent case Omit
Nalt = 4	2	0.9001	0.9019	0.9886	0.5395	0.4136	0	0
	4	0.8998	0.8878	0.9840	−0.2811	−0.4588	1	1
	6	0.8506	0.8208	0.9732	−0.8131	−1.0484	0	0
	8	0.8013	0.7865	0.9501	−1.1599	−1.4624	0	0
Nalt = 7	2	0.9001	0.9028	0.9880	0.5395	0.3694	0	0
	4	0.8998	0.8856	0.9822	−0.2811	−0.5318	1	2
	6	0.8506	0.8166	0.9701	−0.8131	−1.1559	0	2
	8	0.8013	0.7728	0.9433	−1.1599	−1.6268	0	0
Ignored	2	0.9001	0.8891	0.9866	0.5395	0.6551	0	57
	4	0.8998	0.8832	0.9834	−0.2811	−0.1543	1	23
	6	0.8506	0.8311	0.9719	−0.8131	−0.7235	0	5
	8	0.8013	0.7843	0.9474	−1.1599	−1.1375	0	0
Incorrect	2	0.9001	0.9035	0.9868	0.5395	0.3120	0	0
	4	0.8998	0.8819	0.9795	−0.2811	−0.6294	1	3
	6	0.8506	0.8095	0.9643	−0.8131	−1.3041	0	5
	8	0.8013	0.7463	0.9260	−1.1599	−1.8646	0	0

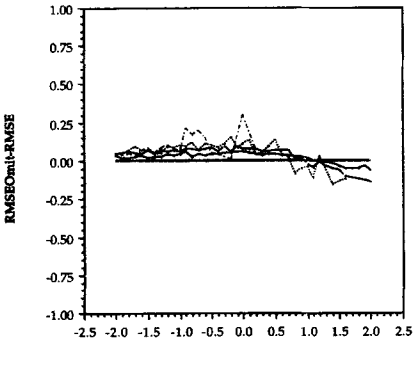
quadrature points by 0.0003 or less. The only exception was for the Nomitted = 6 level, in which the fidelity coefficients differed by 0.001. Doubling the number of quadrature points did not have a meaningful effect on the EAP fidelity coefficients, therefore the use of 20 quadrature points for estimation with the Incorrect condition was not examined. As was the case for the Biweight estimation, the $r_{\theta\hat{\theta}_c}$s were greater than the $r_{\theta\hat{\theta}_o}$s for corresponding Nomitted levels. Although the correlation between the $\hat{\theta}_c$ and $\hat{\theta}_o$ showed the same pattern as was seen with the Biweight estimates, all the EAP correlations were greater than 0.96. Moreover, the comparison of the Ignore Omits and Omits Treated as Incorrect levels revealed the same pattern seen with the Biweight estimates.

Of the three estimation methods, MLE showed the lowest fidelity coefficients for both $\hat{\theta}_c$ and $\hat{\theta}_o$ for corresponding levels of omission (Table 4). As was the case with the Biweight and EAP $\hat{\theta}$s the fidelity coefficients decreased as the number of omits increased, although the difference between the largest and smallest value for a given Nalt level was larger than that seen with the other ability estimation methods (as large as 0.1572 in one condition). In general, for a given Nomitted level, the lowest $r_{\theta\hat{\theta}_o}$s were seen when omitted responses were ignored when estimating the simulee's ability. This also was the condition in which the largest number of nonconvergent cases were observed, although proportionally the nonconvergent cases represented less than 1% of the cases estimated.

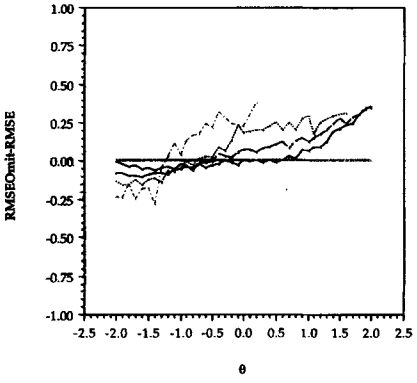
The accuracy of estimation was examined graphically. For all figures only data points based on 10 or more cases were plotted. These figures contain the difference between *RMSE* based on the omission vector [$RMSE(\hat{\theta}_o)$] and *RMSE* based on the corresponding complete vector [$RMSE(\hat{\theta}_c)$] for given level of Nomitted. If for a



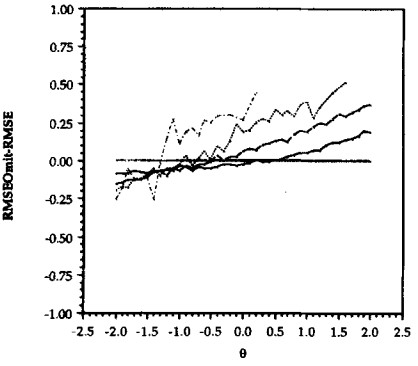
(a) Biweight, RMSE, C=2, Ignore Omits



(b) Biweight, RMSE, C=10, Ignore Omits



(c) Biweight, RMSE, C=2, Omits Treated as Incorrect



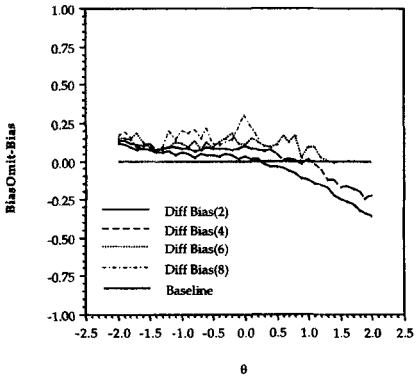
(d) Biweight, RMSE, C=10, Omits Treated as Incorrect

FIGURE 1. RMSE *for Biweight ability estimation*

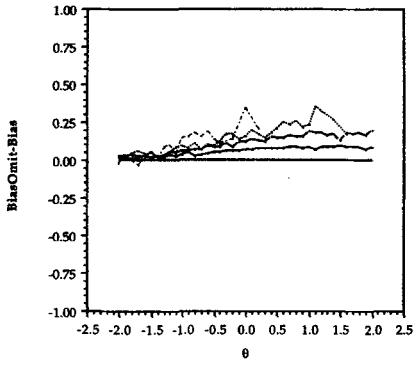
given level of Nomitted, *RMSE* based on the omission vector was the same as *RMSE* based on the corresponding complete vector [(i.e., $RMSE(\hat{\theta}_o) = RMSE(\hat{\theta}_c)$ across the θ scale], then their difference would be 0 and the corresponding line would be equivalent to a baseline of 0. This is true for all the RMSE plots discussed, and the Bias plots have a similar interpretation.

Figure 1 contains *RMSE* as a function of θ. In general, increasing levels of omission led to larger discrepancies between $RMSE(\hat{\theta}_c)$ and $RMSE(\hat{\theta}_o)$ than did lower levels of omission. The effect of omissions was primarily an increase in *RMSE* relative to that obtained with complete response vectors. The most erratic pattern observed corresponded to cases with eight omitted responses, and even for these cases this occurred over a limited θ range. Because the maximum location parameter was 1.29, the patterns displayed above this location may be somewhat idiosyncratic.

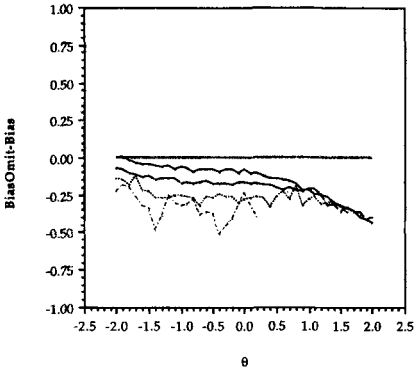
Figure 1b contains the Biweight $C = 10$ *RMSE* results (the $C = 4$, $C = 6$, and $C = 8$ conditions fell predictably between this figure and Figure 1a). This condition represents the least amount of trimming. However, except for slight increases in *RMSE* ($\hat{\theta}_o$) for the eight omitted responses, there appeared to be only a slight effect due to reducing the amount of trimming for simulees located below approximately 0.5. Figure 1c shows the same trim level as Figure 1a, but for the Omits Treated as Incorrect condition. The effect of treating omits as incorrect is revealed in the greater discrepancies between *RMSE* ($\hat{\theta}_o$) and *RMSE*($\hat{\theta}_c$) for simulees throughout the θ range. Simulees at the lower end of the θ continuum were better estimated when treating the omits as incorrect than when complete data were available. However, above approximately $\theta > -1.0$ the increased number of incorrect responses led to poorer estimation than was seen with $\hat{\theta}_c$. This pattern was not evident from the fidelity coefficients and was exhibited for all trim levels (cf. Figure 1d). When omits were treated as incorrect, more pronounced differences



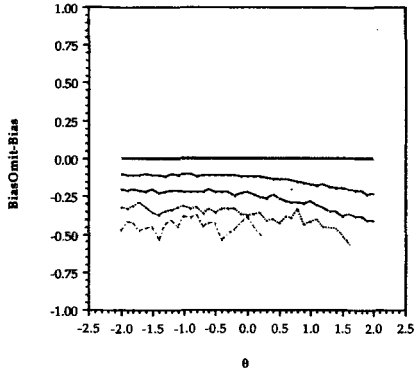
(a) Biweight, Bias, C=2, Ignore Omits



(b) Biweight, Bias, C=10, Ignore Omits

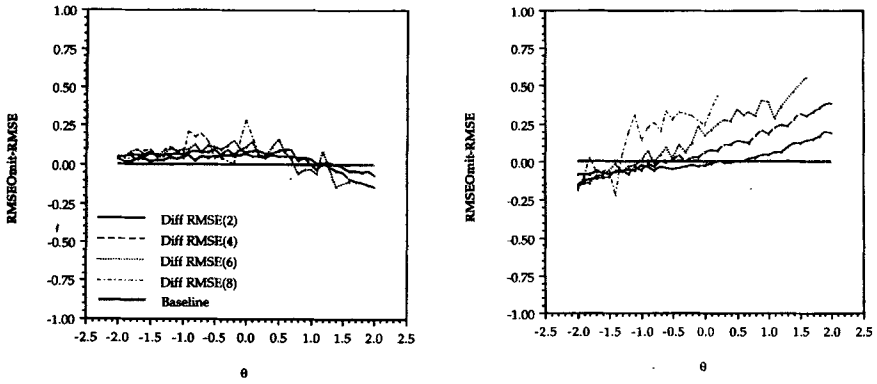


(c) Biweight, Bias, C=2, Omits Treated as Incorrect



(d) Biweight, Bias, C=10, Omits Treated as Incorrect

FIGURE 2. *Bias for Biweight ability estimation*



(a) EAP, RMSE, 10 Quadrature Points, Ignore Omits

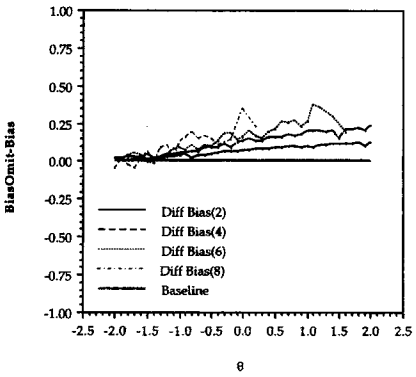
(b) EAP, RMSE, 10 Quadrature Points, Omits Treated as Incorrect

FIGURE 3. RMSE *for EAP ability estimation*

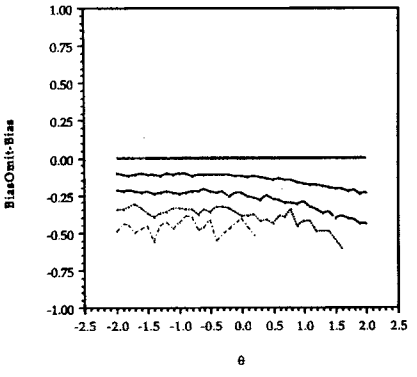
across all trim levels were observed. Specifically, at the higher trim levels (e.g., $C = 2$) the discrepancies across Nomitted levels were less than those at the lower trim levels (e.g., $C = 10$).

The corresponding Biweight Bias plots are presented in Figure 2. Similar to the *RMSE* plots, the Bias figures contain the difference between Bias based on the omission vector [Bias($\hat{\theta}_o$)] and Bias based on the corresponding complete vector [Bias($\hat{\theta}_c$)] for a given level of Nomitted. In general, for the $C = 2$ condition (Figure 2a), increasing omission levels led to increasing levels of bias in $\hat{\theta}_o$. This pattern also can be observed in the $C = 10$ (Figure 2b) condition (the $C = 4$, $C = 6$, and $C = 8$ conditions fell predictably between this figure and Figure 2a). All figures showed a pattern of increasing Bias($\hat{\theta}_o$) as Nomitted increased. The negative differences observed indicate a situation in which there was less bias in $\hat{\theta}_o$ than in $\hat{\theta}_c$, although these $\hat{\theta}_o$s may be somewhat less stable than those below $\theta = 1.29$ (Figure 2a). A comparison of corresponding trim levels across the Ignore Omits/Omits Treated as Incorrect conditions showed an increased level of bias throughout the θ range as well as an "opposite" pattern of bias. The increase in the number of incorrect responses reduced the Bias($\hat{\theta}_o$) relative to Bias($\hat{\theta}_c$). As was observed with *RMSE* when omits were treated as incorrect, the trim level had a more visible effect on the dispersion of the Bias($\hat{\theta}_o$) − Bias($\hat{\theta}_c$) across Nomitted levels. Specifically, less dispersion was seen at higher trim levels than at lower trim levels.

Figures 3 and 4 contain the EAP *RMSE* and Bias results for the 10 quadrature point condition (the results for the 20 quadrature point level were very similar). As with Biweight estimation, increasing levels of omission led to slightly larger *RMSE*($\hat{\theta}_o$); however, this effect of omission at the lower end of the continuum was not very large. The most erratic pattern observed again corresponded to the eight omitted responses condition. In general, the fewer the number of omissions, the more similar *RMSE*($\hat{\theta}_c$) and *RMSE*($\hat{\theta}_o$) were. A comparison of Figure 3a (Ignore Omits) with Figure 3b (Omits Treated as Incorrect) showed a marked discrepancy



(a) EAP, Bias, 10 Quadrature Points, Ignore Omits



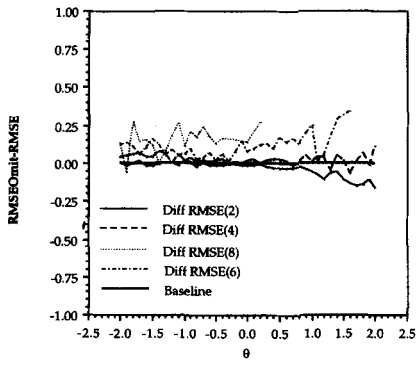
(b) EAP, Bias, 10 Quadrature Points, Omits Treated as Incorrect

FIGURE 4. *Bias for EAP ability estimation*

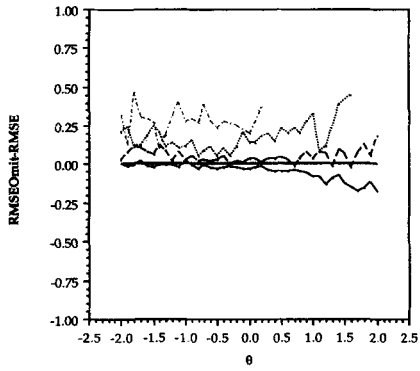
between $RMSE(\hat{\theta}_c)$ and $RMSE(\hat{\theta}_o)$ particularly for the higher omission levels than what had been observed when omits were ignored.

As one would expect from EAP, an examination of the $\hat{\theta}_c$s showed a tendency to overestimate low θs and underestimate at upper proficiency levels. Unlike the Biweight condition, EAP Bias($\hat{\theta}_o$) appeared to increase around $\theta = 0$ for all levels of Nomitted, although the pattern of increasing Bias($\hat{\theta}_o$) as Nomitted increased was still evident above approximately $\theta = -1.0$. Bias($\hat{\theta}_c$) and Bias($\hat{\theta}_o$) were virtually identical at the lowest end of the θ continuum when omits were ignored (Figure 4a). For the Omits Treated as Incorrect condition (Figure 4b), it appeared that the EAP $\hat{\theta}_o$s were less biased, on average, then $\hat{\theta}_c$s were. However, inspection of Bias($\hat{\theta}_o$) and Bias($\hat{\theta}_c$), rather than their difference, showed Bias ($\hat{\theta}_o$) as a function of θ exhibited the same pattern as Bias($\hat{\theta}_c$), but with the point at which bias = 0 shifted down the proficiency scale about 0.8 to 1.3 logits depending on the omission level. In fact, Bias($\hat{\theta}_o$) showed a progressively increasing underestimation of θ as θ increased so that $|\text{ Bias}(\hat{\theta}_o)| > |\text{ Bias}(\hat{\theta}_c)|$ when $\theta > 0.5$ for the Nomitted = 2; for the other levels of omit this crossover point was −0.4 (Nomitted = 4), −0.7 (Nomitted = 6), and −1.3 (Nomitted = 8). Therefore, the pattern in Figure 4b is a function, in part, of subtracting numbers with opposite signs at some points along the θ scale and at other points subtracting negative numbers from one another.

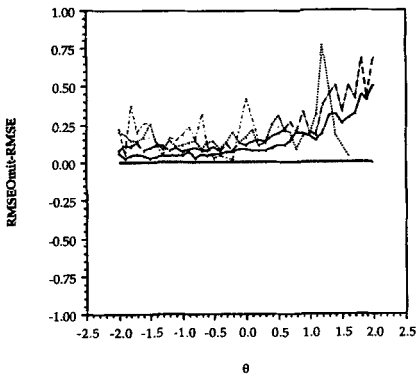
Figure 5 contains the *RMSE* plots for MLE ability estimation. The pattern of increasing $RMSE(\hat{\theta}_o)$ as a function of increasing Nomitted also was observed with MLE. Although Figures 5a, 5b, and 5d show that for the Nomitted = 2 condition $RMSE(\hat{\theta}_o)$ was less than $RMSE(\hat{\theta}_c)$, the difference was relatively small and may be attributed to random sampling fluctuations; *t* tests on the ln(*RMSE*) showed that there were no statistically significant ($\alpha = 0.05$) differences between *RMSE* ($\hat{\theta}_o$) and $RMSE(\hat{\theta}_c)$ for Nalt levels of 4 and 7 and for Omits Treated as Incorrect. The effect of the number of omits on the accuracy of $\hat{\theta}_o$ was more pronounced with MLE than with either Biweight or EAP. It appeared that MLE ability estimation



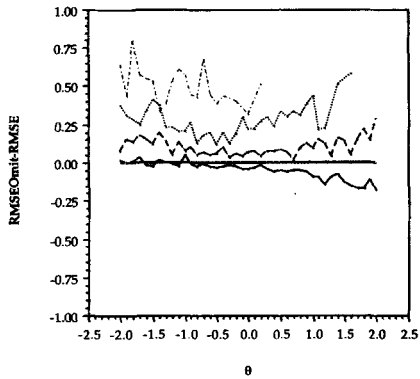
(a) MLE, RMSE, Number of Alternatives=4



(b) MLE, RMSE, Number of Alternatives=7



(c) MLE, RMSE, Ignore Omits

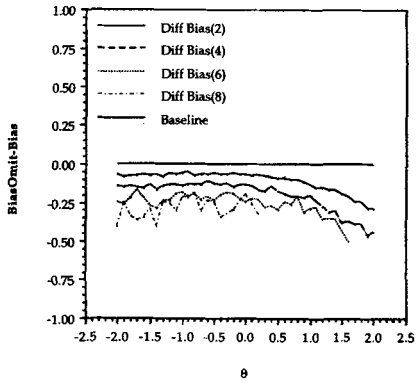


(d) MLE, RMSE, Omits Treated as Incorrect

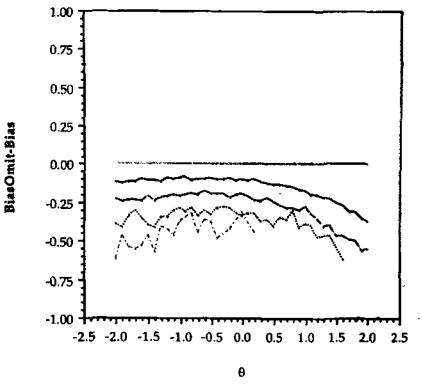
FIGURE 5. RMSE *for MLE ability estimation*

was not very affected by two or four omissions (10.3% or fewer omits), but omitting more than four items had a marked increase in $RMSE(\hat{\theta}_o)$ (Figures 5a and 5b). Comparing Figures 5a and 5b shows that increasing the number of alternatives from four to seven led to increases in $RMSE(\hat{\theta}_o)$ for Nomitted = 4, 6, and 8 conditions. Ignoring omits in estimating proficiency (Figure 5c) decreased the accuracy of $\hat{\theta}_o$ above $\theta = -0.5$, but below this point the differences between $RMSE(\hat{\theta}_o)$ and $RMSE(\hat{\theta}_c)$ were similar to those observed when the number of alternatives was four. Figure 5d shows that treating the omitted responses as incorrect led to the largest discrepancies between $RMSE(\hat{\theta}_o)$ and $RMSE(\hat{\theta}_c)$ of all conditions investigated. This discrepancy became more pronounced as the number of omissions increased.

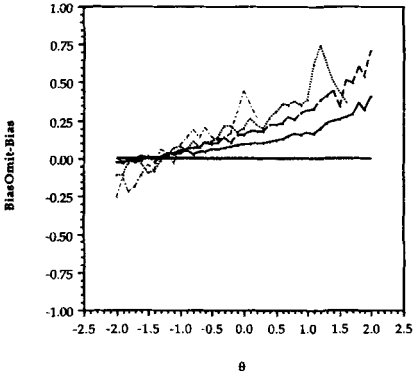
The corresponding Bias plots are presented in Figure 6. Unlike Biweight and EAP, there was, on average, a relatively consistent positive Bias $(\hat{\theta}_c)$ throughout the proficiency continuum for MLE. Except for the Ignoring Omits condition, the relationship of Bias$(\hat{\theta}_o)$ and Bias$(\hat{\theta}_c)$ showed a pattern similar to that seen with the



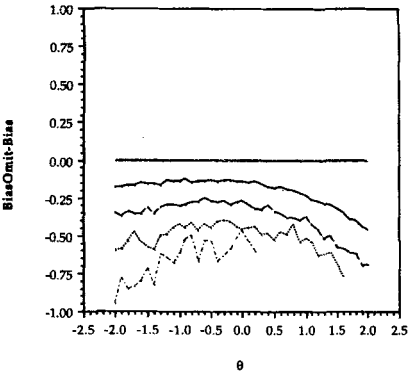
(a) MLE, Bias, Number of Alternatives=4



(b) MLE, Bias, Number of Alternatives=7



(c) MLE, Bias, Ignore Omits



(d) MLE, Bias, Omits Treated as Incorrect

FIGURE 6. *Bias for MLE ability estimation*

EAP Omits Treated as Incorrect condition (Figure 4b). Specifically, for omission vectors with two omits, Lord's approach led to less positively biased $\hat{\theta}_o$ than would normally be observed with complete vectors for part of the θ scale. This also was true for treating omits as incorrect. However, inspection of the Bias($\hat{\theta}_o$)s for the 4, 6, and 8 Nomitted levels showed increasing negatively biased $\hat{\theta}_o$s as a direct function of Nomitted. The particular pattern displayed for the 4, 6, and 8 Nomitted levels in these figures was a result of subtracting Bias($\hat{\theta}_c$) from negative Bias($\hat{\theta}_o$) where | Bias($\hat{\theta}_o$) | > Bias($\hat{\theta}_c$). Moreover, for the 4, 6, and 8 Nomitted levels, the largest negatively biased $\hat{\theta}_o$s were found when omits were treated as incorrect and the smallest when specifying four alternatives. When omits were ignored in estimating $\hat{\theta}$ (Figure 6c) and for the Nomitted = 2 and 4 levels, $\hat{\theta}_o$ was more positively biased than $\hat{\theta}_c$ for almost the entire θ range. With Nomitted = 6, there was a negative bias in $\hat{\theta}_o$ at lower θs [Bias($\hat{\theta}_c$) was positive in this range] that became a positive bias as θ increased [corresponding Bias($\hat{\theta}_c$) became negative as θ in-

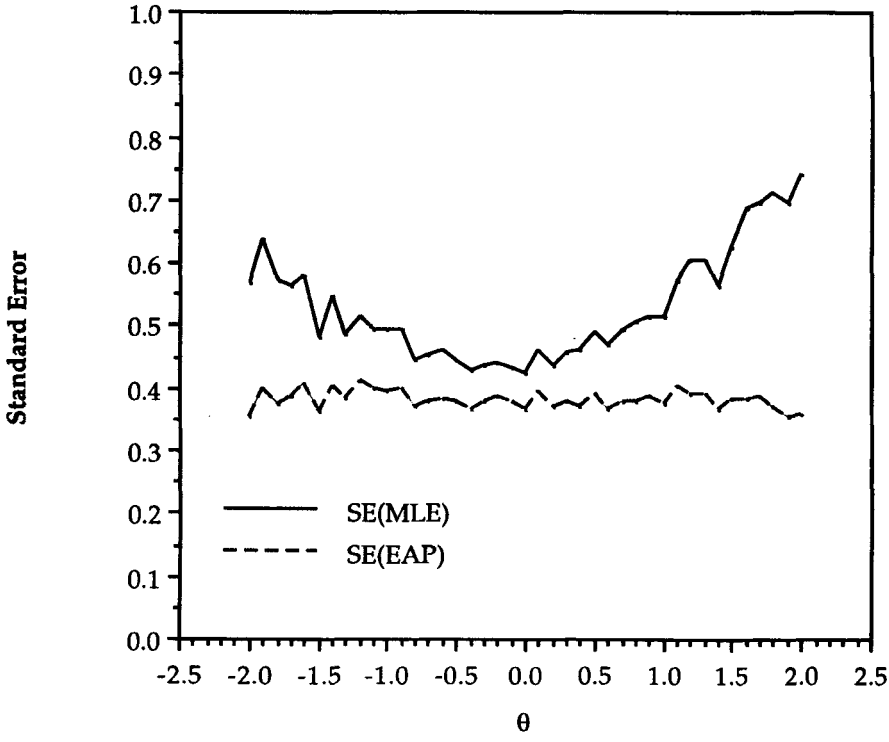


FIGURE 7. *Average standard error for EAP and MLE for complete vectors*

creased]. For Nomitted = 8, Bias($\hat{\theta}_o$) and the corresponding Bias($\hat{\theta}_c$), in general, were negatively biased throughout the proficiency continuum, although $\hat{\theta}_o$ showed greater negative bias.

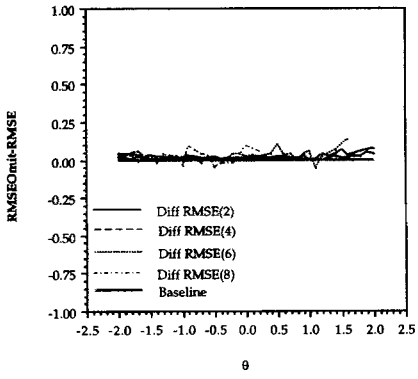
Phase 2

Given that the *RMSE* and Bias results for Nalt = 4 were better than that for Nalt = 7, it was hypothesized that specifying the number of alternatives as two might further reduce *RMSE*($\hat{\theta}_o$) and Bias($\hat{\theta}_o$). Additional justification for the choice of Nalt = 2 was discussed above. Although EAP showed *RMSE*($\hat{\theta}_o$) that were less than that of MLE, MLE showed less bias at the ends of the proficiency continuum than did EAP. To determine whether EAP or MLE was performing better overall, the variance error of estimate [VEE(θ)] was calculated:

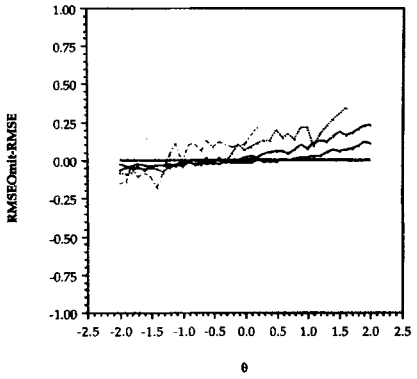
$$\text{VEE}(\theta) = RMSE(\theta)^2 - \text{Bias}(\theta)^2. \quad (13)$$

The square root of VEE is presented in Figure 7. As one can see, EAP performed better than MLE across the θ continuum.

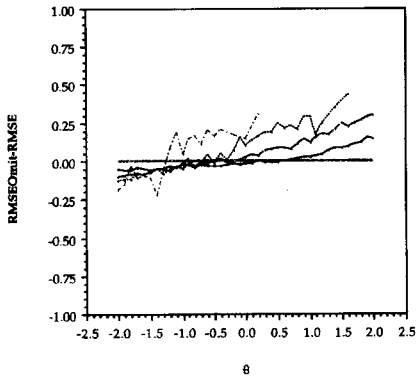
Specifying that the number of alternatives was two and using 10 quadrature points, EAP $\hat{\theta}$s were obtained for each level of Nomitted. The corresponding



(a) EAP, RMSE, Number of Alternatives=2



(b) EAP, RMSE, Number of Alternatives=4



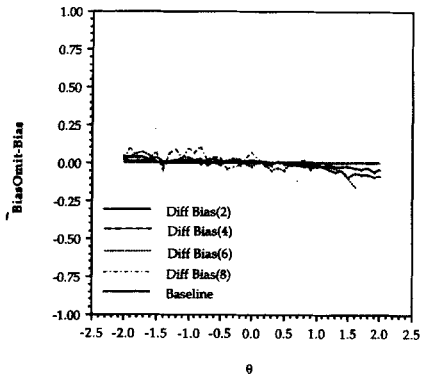
(c) EAP, RMSE, Number of Alternatives=7

FIGURE 8. RMSE *for EAP ability estimation using Nalt*

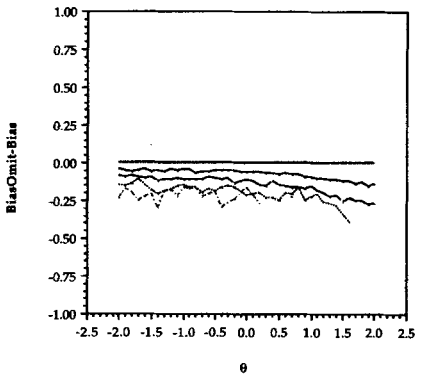
RMSE($\hat{\theta}_o$)s and Bias($\hat{\theta}_o$)s are presented in Figure 8 (above) and Figure 9, respectively. Figure 8a shows that the agreement between *RMSE*($\hat{\theta}_o$) and *RMSE*($\hat{\theta}_c$) increased relative to what had been observed in Figure 3 for all Nomitted levels. The use of Nalt values of 4 and 7 paralleled the findings with MLE in that the larger the Nalt value the worse the recovery of the θ. Similarly, the discrepancy between Bias($\hat{\theta}_o$) and Bias($\hat{\theta}_c$) also was reduced when a Nalt of 2 was used (cf. Figure 4). Increasing Nalt to 7 increased the discrepancy between Bias($\hat{\theta}_o$) and Bias($\hat{\theta}_c$). Of the three values of Nalt examined, the Nalt of 2 resulted in the best estimation of θ relative to that found with complete data.

Discussion

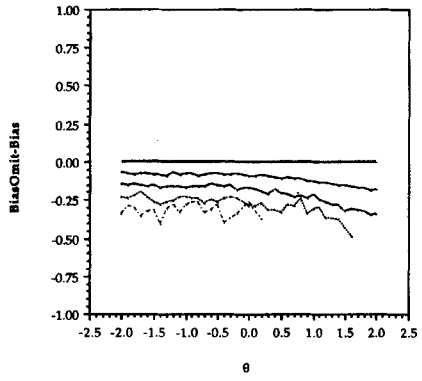
The above results seemed to indicate that omits should not be treated as incorrect. As would be expected, the accuracy of $\hat{\theta}$ decreased as the number of omissions increased, particularly for examinees located toward the upper end of the



(a) EAP, Bias, Number of Alternatives=2



(b) EAP, Bias, Number of Alternatives=4



(c) EAP, Bias, Number of Alternatives=7

FIGURE 9. *Bias for EAP ability estimation using Nalt*

proficiency scale. Although ignoring "nonignorable" omits affected all ability estimation methods, it appeared that ignoring omits had a greater impact on $\hat{\theta}$s under MLE than with other methods (cf. Figures 1a, 3a, and 5c). Contrasting the results for when omits were treated as incorrect with those when omits were ignored appeared to indicate that more accurate ability estimates (e.g., in terms of *RMSE*, Bias, $\bar{\hat{\theta}}_c - \bar{\hat{\theta}}_o$) are obtained throughout the proficiency continuum when omits are ignored rather than when scored as incorrect, although the linear relationship between θs and $\bar{\hat{\theta}}_o$s was not particularly affected by whether the omits were treated as incorrect or ignored.

For Biweight ability estimation and regardless of whether omits were ignored or treated as incorrect responses, the trim level appeared to have only a slight effect on $r_{\theta\hat{\theta}_o}$. When ignoring omits, the interaction between trim level, number of omits, and θ made it difficult to recommend a particular trim level over another (see Figures 1a, 1b, 2a, 2b). In terms of $\bar{\hat{\theta}}_c - \bar{\hat{\theta}}_o$, increasing Nomitted led to a greater $\bar{\hat{\theta}}_c - \bar{\hat{\theta}}_o$

discrepancy for all trim levels but was most pronounced under heavy trimming (across Nomitted and trim levels the average discrepancy was 0.0778). In short, the level of trimming employed under this condition depends on what is considered most important. In contrast, when omits are treated as incorrect, it appears that the heavy trim level ($C = 2$) is preferable to less trimming.

In EAP ability estimation, the use of 20 quadrature points appeared to have no substantial effect on improving the statistics over using 10 quadrature points. Unless one is concerned with the accuracy of standard errors (De Ayala, Schafer, & Sava-Bolesta, 1995) there appears to be no compelling reason to use 20 quadrature points.

These results appear to suggest a possible strategy for handling omits by the IRT practitioner. The use of EAP where omits are replaced by 0.5 (i.e., Nalt = 2) not by the reciprocal of an item's number of alternatives, is an estimation method that should be considered; this also would most likely be true for MLE. In this case, Nalt is a misnomer and should not be interpreted to indicate the use of two-alternative items. Moreover, this strategy does not assume that all examinees would answer an item (instead of omitting it) if their chances of correctly answering it were greater than 1/Nalt (i.e., the random guessing value). Specifying Nalt = 2 simply minimizes the magnitude of the possible discrepancy between the expected (using random guessing model) and the predicted probability of a correct response based on an IRT model. As such, one is simply imputing a "response" for a binomial variable and thereby "smoothing" irregularities in the likelihood function. This could easily be done in BILOG by specifying Nalt as 2.

It should be noted that for this study the data were generated, in part, according to a 3-PL model. To the extent that a model with a pseudo-guessing parameter more accurately describes the data than a model without this feature, then the use of models without the pseudo-guessing parameter may not produce results comparable to those seen here.

Notes

We would like to thank the Editor and two anonymous reviewers for suggestions and comments that improved the quality of this manuscript.

A version of this article that examined the use of the nominal response model with omitted responses can be obtained from ERIC Document Reproduction Service (No. TM 032009), from the first author, or from the first author's website at http://www.tc.unl.edu/rdeayala/RJsite.html.

References

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

De Ayala, R. J., Schafer, W. D., & Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *British Journal of Mathematical and Statistical Psychology, 48*, 385–405.

Little, R., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247–264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725–737.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring: With binary logistic model.* Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Wu, P. (1988). *Inferring examinee ability when some item responses are missing* (RR 88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

Stocking, M. L., Eignor, D., & Cook, L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed and true score equating procedures* (RR-88-41). Princeton, NJ: Educational Testing Service.

Thissen, D. J. (1991). MULTILOG (Version 6.0). Scientific Software, Inc. Mooresville, IN.

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

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

Authors

R. J. DE AYALA is a professor and chairperson of the Department of Educational Psychology at the University of Nebraska-Lincoln, Teachers College Hall, Room 21, Lincoln, NE 68588. His research interests include psychometrics and applied statistics.

BARBARA S. PLAKE is a professor and director of the Buros Center for Testing at the University of Nebraska-Lincoln, Teachers College Hall, Room 21, Lincoln, NE 68588. Her research interests include applied educational measurement.

JAMES C. IMPARA is an associate professor and director of the Buros Institute for Assessment Consultation and Outreach at the University of Nebraska-Lincoln, Teachers College Hall, Room 21, Lincoln, NE 68588. His research interests include applied educational measurement.