LSAC RESEARCH REPORT SERIES

Violations of Ignorability in Computerized Adaptive Testing

Cees A. W. Glas
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 04-04
September 2006

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

In a computerized adaptive test (CAT), each test taker responds to a limited subset of items from the available item bank, as test items are chosen throughout the testing session to match the most current estimate of the test taker's ability level. The statistical theory of computerized adaptive testing is based on the assumption that the dependence between item selection and ability estimation does not result in any bias of the ability estimate. This assumption holds because, in general, the missing-data process is ignorable, that is, item selection is dependent only on the test taker's responses to the items he or she has been administered, and the remaining items in the item pool may be ignored. There are at least two cases for which this assumption does not hold.

1. If item selection is also based on an impressionistic estimate of the test taker's ability, or on some other covariates that are not explicitly modeled. This case occurs if the first item in the CAT is chosen based on some prior information about the test taker or if the CAT is stratified for the purpose of content balancing.

2. If item review is allowed and test takers change their responses to previous items. If test takers change earlier responses, the item-selection design is no longer completely determined by the observed responses.

The violation of the ignorability assumption for the missing-data process does not automatically lead to bias. In this report, the following two different situations have been examined:

1. Estimation of the ability parameters based on auxiliary information about the test taker's ability and the allowance of item review. Both analytically and through simulation studies, it was shown that this case of violation of the ignorability assumption did not lead to a gross inflation of bias.

2. Calibration of item and population parameters using a method known as maximum marginal likelihood estimation. Through simulation studies, it was shown that this case of violation of the ignorability assumption did result in bias. An analytical explanation of the result is given.

Abstract

Using auxiliary information and allowing item review in computerized adaptive testing produces a violation of the ignorability principle for missing data (Rubin, 1976) that may bias parameter estimates in IRT models. However, the violation of ignorability does not automatically lead to bias. In this report, two situations are distinguished.

1. Estimation of the proficiency parameters in computerized adaptive testing using auxiliary information about proficiency and allowing item review, where the item parameters are considered known. Both analytically and through simulation studies, it is shown that the violation of ignorability does not lead to a gross inflation of bias.

2. Calibration of item and population parameters using maximum marginal likelihood estimation. Through simulation studies it is shown that violation of ignorability does result in bias. An analytical explanation of the result is given.

Introduction

In a computerized adaptive test (CAT), a student responds to a limited subset of items sampled from an item bank. The statistical theory of CAT is based on the fact that, in general, the dependence between the item selection procedure and the unobserved proficiency level does not result in bias of the proficiency estimate. This is because, again in general, the missing data process is ignorable, because the item administration design is completely determined by observed responses. There are at least two situations where this no longer holds.

1. If the choice of the items in the test is not only based on the responses, but also on an independent estimate of the student's proficiency, or on some other covariates that are not explicitly modeled. This may be especially eminent for the choice of first items in the CAT.

2. If item review is allowed, that is, if the students are allowed to change their responses to previous items. In this case, the design is no longer completely determined by the responses alone, because the responses have changed.

Ignorability

In applications of IRT to CAT, students seldom respond to all available items. Every student is administered a virtually unique test by the very nature of the item selection mechanism of CAT. For every student, indexed n, the test administration design can be described by a test administration vector d_n, with elements d_{ni}, $i = 1, ..., I$, where I is the number of items in the item pool. The item administration variable d_{ij} is equal to one if student n responded to item i, and zero otherwise. The design for all students is represented by an $N \times I$ design matrix D. The response variable x_{nj} is defined to have a value of one if a correct response is observed, zero if an incorrect response is observed, and equal to an arbitrary constant if no response is observed.

In the context of CAT, it is an interesting question as to whether estimates can be calculated treating the design as fixed and maximizing the likelihood of the parameters conditional on D. If so, the design is called ignorable (Rubin, 1976). Using Rubin's theory on ignorability of designs, this question is extensively studied by Mislevy and Wu (1996). They conclude that the administration design is ignorable in adaptive testing. Further, their conclusions also have consequences for the calibration stage, because the design is usually also ignorable in estimation using data from tests targeted to the proficiency level of the students.

In the present report, we assess a number of situations where ignorability is violated. Therefore, first the ignorability principle will be outlined in some detail. As explained above, let D be a missing data indicator, and let the potential responses be partitioned into the actually observed responses x_{obs} and the unobserved responses x_{mis}. The parameter of interest is denoted by ξ, and the probability model for x_{mis} depends on parameters ϕ. The key concept in the theory of ignorability is *missing at random* (MAR), which holds if

$$p(D \mid x_{obs}, x_{mis}, \phi, y) = p(D \mid x_{obs}, \phi, y),$$

where y are covariates that might play a role. So MAR holds if the missing data indicators D do not depend on the missing data x_{mis}, in fact, they only depend on the observed data x_{obs}, and possibly on covariates y. Then, there is a technical condition. In a frequentist framework, the condition is that ϕ and θ are distinct, that is, the space of ϕ and θ factorizes into a ϕ-space and a θ-space, and the two sets of parameters have no mutual functional restrictions. In a Bayesian framework, ϕ and θ are distinct if $p(\phi \mid \theta, y) = p(\phi \mid y)$, that is, if they have independent priors. Rubin (1976) proved the following.

Theorem

If ϕ and θ are distinct, and MAR holds, then in a frequentist framework $p(x_{obs}, D \mid \phi, y) \propto p(x_{obs}, \mid \theta, y)$ and in a Bayesian framework $p(\theta \mid x_{obs}, D, y) \propto p(\theta \mid x_{obs}, y)$. The frequentist version implies that inferences such as maximum likelihood estimation can be based on the likelihood of the observed data, $p(x_{obs}, \mid \theta, y)$, and the process causing the missingness does not have to be taken into account. In the same manner, the Bayesian version implies that inferences can be based on a posterior $p(\theta \mid x_{obs}, y)$ that ignores the probability model for D. It should be noted that conditioning on D may produce an overestimate of the sample variability of the data and, consequently, an underestimate of the standard error of the estimate of θ. Unbiased inferences on standard errors might be obtained if the data are also *observed at random*, that is, if $p(D \mid x_{obs}, x_{mis}, \phi, y) = p(D \mid x_{mis}, \phi, y)$, so x_{obs} does not depend on D.

Ignorability in CAT directly follows from the theorem: in CAT the item selection process completely depends on the observed responses and is completely independent of the unobserved responses. Further, ignorability also holds when CAT data are used to calibrate the item and population parameters using maximum marginal likelihood (MML; see Bock & Aitkin, 1981; the impact of targeted designs on MML estimation was studied by Glas, 1988; Mislevy & Sheehan, 1989; Eggen & Verhelst, 1992; and Mislevy & Chang, 2000).

In the present report, two cases are investigated where the observed data no longer determine the design D, the case where auxiliary information on the students' proficiency is used to select items and the case of item review where the original responses are no longer available. The impact of these two cases will be studied for the estimation of θ in the operational CAT phase, and the estimation of the item parameters using CAT data in the calibration phase.

Estimation of θ in CAT

The consistency of the estimates of θ in CAT can be established using the ignorability principle: the observed responses completely determine the choice of the items. Two situations will be studied where the item choice is no longer determined by the observed response, those situations being item review and item choice based on auxiliary information.

Item Review

In item review, students can alter earlier given responses. It is assumed here that the original responses governing the choice of the items are not recorded, so that ignorability no longer holds. Let x_{obs} stand for the available response pattern, and let x_{mis} stand for the original pattern determining the item administration design d. The item parameters, say β, are considered known, and θ is the estimand. In principle, inferences should be based on

$$p(x_{obs}, x_{mis}, d; \theta, \beta) = p(x_{obs}; \theta, \beta)p(x_{mis}; \theta, \beta)p(d \mid x_{mis}; \theta, \beta)$$

$$= p(x_{obs}; \theta, \beta)p(x_{mis}; \theta, \beta),$$

where the second equation holds because $p(d \mid x_{mis}; \theta, \beta) = 1$. For further clarification, we take the logarithm of both sides of the equation to obtain

$$\log p(x_{obs}, x_{mis}, d; \theta, \beta) = \log p(x_{obs}; \theta, \beta) + \log p(x_{mis}; \theta, \beta).$$

If x_{mis} were available, from the ignorability principle for CAT it follows that both maximization of $\log p(x_{obs}, x_{mis}, d; \theta, \beta)$ and $\log p(x_{mis}; \theta, \beta)$ would lead to consistent estimates of θ, the latter estimates being less efficient of course. This, in turn, has the consequence that maximization of $\log p(x_{obs}; \theta, \beta)$ produces a consistent estimate of θ.

Auxiliary Information

An analogous argument holds when auxiliary information on θ is available. Here, we only consider the case where this information can be summarized as an estimate θ_0 that can be viewed as a consistent estimate of θ. More specifically, assume that θ_0 has a normal distribution with mean θ and standard deviation σ. Further, part of the design d is targeted at θ_0 rather than at the running estimate of θ inferred from x_{obs}. Then the complete log-likelihood decomposes as

$$\log p(x_{obs}, \theta_0, d; \theta, \beta) = \log p(x_{obs}; \theta, \beta) + \log p(\theta_0; \theta, \sigma).$$

As above, the consistency of the estimates based on $\log p(x_{obs}, \theta_0, d; \theta, \beta)$ and $log\ p(\theta_0; \theta, \sigma)$ assures the consistency of the estimates based on $\log p(x_{obs}; \theta, \beta)$.

Simulation Studies

A number of simulation studies were conducted to elucidate the two cases discussed above. The one-parameter logistic model (1-PLM) was used to avoid contamination of the results by the possibly poor identification of the two-parameter logistic model (2-PLM) and the three-parameter logistic model (3-PLM). The proficiency parameter assumed five values (θ = –2.0, –1.0, 0.0, 1.0, 2.0), and for every value the responses of 10,000 simulees were generated. Unless indicated otherwise, the starting value of the proficiency estimate was equal to zero. The test length was either 20 or 40. To avoid possible affects of the composition of the item bank, it was assumed that the optimal item was always available. So the size of the item bank was assumed infinite. The proficiency parameter was estimated by maximum likelihood, and maximum information was used as a selection criterion. The following eight conditions were introduced.

1. Random item selection. In this condition, for every simulee a new set of item parameters were randomly drawn from the standard normal distribution, and responses to this randomly assembled test were generated. This condition did not entail CAT; it was used as a base line for reference.

2. Computerized adaptive testing.

3. CAT with item review. In this condition, new responses were generated for all the selected items, so the condition is far more extreme than what can be expected in real life testing situations.

4. CAT with item review only for proficiency levels $\theta > 0.0$. In this first set of simulations, the results will just be a combination of the two previous conditions; the purpose of this condition will become apparent in the simulation studies pertaining to item calibration.

5. CAT with the first half of the test items chosen to be optimal at the true proficiency value.

6. CAT with all items chosen to be optimal at the true proficiency value.

7. CAT with the first half of the test items chosen to be optimal at θ_0, where θ_0 was drawn from a normal distribution with a mean equal to the true proficiency parameter, and with a standard deviation equal to 1.0.

8. CAT with the first half of the test items chosen to be optimal at θ_0, where θ_0 was drawn from a normal distribution with a mean equal to the true proficiency parameter, and with a standard deviation equal to 2.0.

The results are shown in Tables 1 and 2 for test lengths of 20 and 40 items, respectively.

TABLE 1
Squared bias and standard errors for θ—20 items per test

Item Selection Mode	θ	Bias	S.E.
Random Selection	–2.0	0.15	0.74
	–1.0	0.11	0.54
	0.0	0.01	0.50
	1.0	0.15	0.52
	2.0	0.26	0.75
CAT	–2.0	0.15	0.70
	–1.0	0.01	0.49
	0.0	0.00	0.44
	1.0	0.02	0.49
	2.0	0.16	0.73
CAT with Item Review	–2.0	0.16	0.71
	–1.0	0.02	0.48
	0.0	0.00	0.46
	1.0	0.01	0.47
	2.0	0.13	0.70
CAT with Item Review if $\theta > 0.0$	–2.0	0.15	0.71
	–1.0	0.11	0.49
	0.0	0.00	0.46
	1.0	0.02	0.47
	2.0	0.14	0.70
50% optimal at true θ	–2.0	0.00	0.48
	–1.0	0.00	0.46
	0.0	0.00	0.46
	1.0	0.00	0.46
	2.0	0.00	0.48
100% optimal at true θ	–2.0	0.02	0.46
	–1.0	0.05	0.45
	0.0	0.10	0.44
	1.0	0.14	0.45
	2.0	0.18	0.46
50% Initial Responses at θ_0 with s.d.(θ_0) = 1.0	–2.0	0.00	0.49
	–1.0	0.00	0.49
	0.0	0.01	0.49
	1.0	0.00	0.49
	2.0	0.00	0.50
50% Initial Responses at θ_0 with s.d.(θ_0) = 2.0	–2.0	0.00	0.53
	–1.0	0.01	0.52
	0.0	0.00	0.52
	1.0	0.01	0.53
	2.0	0.00	0.53

TABLE 2
Squared bias and standard errors for θ—40 items per test

Item Selection Mode	θ	Bias	S.E.
Random Selection	–2.0	0.08	0.51
	–1.0	0.06	0.38
	0.0	0.06	0.34
	1.0	0.08	0.36
	2.0	0.12	0.43
CAT	–2.0	0.01	0.39
	–1.0	0.00	0.33
	0.0	0.00	0.32
	1.0	0.00	0.33
	2.0	0.00	0.40
CAT with Item Review	–2.0	0.02	0.36
	–1.0	0.02	0.33
	0.0	0.05	0.32
	1.0	0.07	0.32
	2.0	0.09	0.35
CAT with Item Review if $\theta > 0.0$	–2.0	0.09	0.40
	–1.0	0.03	0.32
	0.0	0.03	0.33
	1.0	0.08	0.37
	2.0	0.09	0.39
50% optimal at true θ	–2.0	0.00	0.32
	–1.0	0.00	0.32
	0.0	0.00	0.33
	1.0	0.00	0.32
	2.0	0.00	0.32
100% optimal at true θ	–2.0	0.02	0.36
	–1.0	0.02	0.33
	0.0	0.05	0.32
	1.0	0.07	0.32
	2.0	0.09	0.35
50% Initial Responses at θ_0 with s.d.(θ_0) = 1.0	–2.0	0.00	0.34
	–1.0	0.00	0.34
	0.0	0.00	0.33
	1.0	0.00	0.33
	2.0	0.00	0.33
50% Initial Responses at θ_0 with s.d.(θ_0) = 2.0	–2.0	0.00	0.36
	–1.0	0.00	0.37
	0.0	0.00	0.36
	1.0	0.00	0.36
	2.0	0.00	0.37

For every replication, the mean squared error was decomposed into the squared bias and the sampling variance. The resulting bias and standard error are shown in the two last columns of Tables 1 and 2. The following conclusions can be drawn.

1. Compared to random item selection, CAT both reduced the bias and the standard error.

2. The pattern and magnitude of bias and standard error were not changed by item review.

3. Choosing the first half of the test optimal at the true θ reduces both the bias and the standard error compared to CAT.

4. Choosing the complete test to be optimal at the true θ reduces the standard error further, but it also slightly inflates the bias.

5. Choosing the initial items optimal at θ_0 leads to reduction of bias and standard error compared to CAT, and standard errors slightly inflate when the variance of θ_0 increases.

Item Calibration Using CAT Data

Marginal maximum likelihood (MML) estimation is probably the most used technique for item calibration. For the 1-PLM, 2-PLM and 3-PLM, the theory was developed by such authors as Bock and Aitkin (1981), Thissen (1982), Rigdon and Tsutakawa (1983), and Mislevy (1984,1986). Computations can be made using the software package BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996).

MML estimation derives its name from maximizing the log-likelihood that is marginalized with respect to θ, rather than maximizing the joint log-likelihood of all person parameters θ and item parameters β. Let η be a vector of all item and population parameters. Then the marginal likelihood of η is given by

$$\log L(\eta; \mathbf{X}, \mathbf{D}) = \sum_n \log \int p(x_n \mid d_n, \theta_n, \beta) g(\theta_n; \lambda) d\theta_n. \tag{1}$$

The reason for maximizing the marginal rather than the joint likelihood is that maximizing the latter does not lead to consistent estimates. This is related to the fact that the number of person parameters grows proportional with the number of observations, and, in general, this leads to inconsistency (Neyman & Scott, 1948). Simulation studies by Wright and Panchapakesan (1969) and Fischer and Scheiblechner (1970) show that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) have shown that marginal maximum likelihood estimates of structural parameters, say the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models. It has been shown that this also holds for data emanating from CAT (Glas, 1988).

MML estimation equations are usually solved using the EM-algorithm (Dempster, Laird, & Rubin, 1977), which is an algorithm for finding the maximum of a likelihood marginalized over unobserved data. The present application fits this framework when the response patterns are viewed as observed data and the proficiency parameters as unobserved data.

Consider the response pattern of one student; the index n is dropped for convenience. In a situation of item review, the contribution to the log-likelihood given the original data x_{mis} and the reviewed data x_{mis} can be written as

$$\log p(x_{obs}, x_{mis}, d; \eta) = \log \int p(x_{obs} \mid d, \theta, \beta) p(x_{mis}, d; \theta, \beta) g(\theta; \lambda) d\theta$$

$$= \log \int p(x_{obs} \mid d, \theta, \beta) p(\theta \mid x_{mis}, d; \beta, \lambda) p(x_{mis}, d; \beta, \lambda) d\theta$$

$$= \log p(x_{mis}, d; \beta, \lambda) + \log \int p(x_{obs} \mid d, \theta, \beta) p(\theta \mid x_{mis}, d; \beta, \lambda) d\theta.$$

Note that this contribution now consists of a term $\log p(x_{mis}, d; \beta, \lambda)$ and a term $\log \int p(x_{obs} \mid d, \theta, \beta) p(\theta \mid x_{mis}, d; \beta, \lambda)$. The former gives rise to a log-likelihood associated with a CAT design. If x_{mis} were observed, these data could be used to obtain consistent estimates of η. The latter term is the expectation of the probability of x_{obs} with respect to the posterior distribution $p(\theta \mid x_{mis}, d; \beta, \lambda)$. However, if the missing data process is ignored, the expectation of $p(x_{obs} \mid d, \theta, \beta)$ is considered with respect to $g(\theta; \lambda)$, that is, the log-likelihood then becomes a sum of terms

$$\log \int p(x_{obs} \mid d, \theta, \beta) g(\theta; \lambda) d\theta. \tag{2}$$

The effect is that $p(x_{obs} \mid d, \theta, \beta)$ is averaged over the wrong proficiency distribution, that is, a distribution with a wrong location parameter and a wrong scale parameter. To assess the effect, consider two students, one with a high θ-value and one with a low θ-value. The first student is administered difficult items, the second student is administered easy items. However, in (2) both their θ-values are assumed to be drawn from the same distribution. As a result, the easy items are overestimated, and the difficult items are underestimated.

The effect is due to ignoring the covariates x_{mis} and d. When the design is governed by auxiliary information about θ, say θ_0, the situation is essentially the same: when the covariate θ_0 is ignored, the proper

posterior $p(\theta \mid \theta_0; \beta, \lambda)$ is replaced with $g(\theta; \lambda)$, and the result is bias in the estimates of η.

Simulation Studies

To assess the magnitude of the bias caused by ignoring covariates, simulation studies were conducted with conditions similar to the conditions of the simulation studies reported above. The difference is that the data are now generated for 1,000 simulees with parameters drawn from the standard normal distribution; the item bank consisted of 200 items equally spaced between –2.0 and 2.0, and the test length was 20 items. In every condition 100 replications were made. In the condition of random item selection, the test of 20 items was re-sampled from the item bank for every simulee. MML estimates of the item parameters were computed under the assumption that θ had a standard normal distribution.

The results are shown in Table 3. For 5 items from the item bank, the three last columns give the bias, standard error, and mean of the estimates over the replications. The following conclusions can be drawn.

1. Compared to random item selection, CAT greatly reduced the standard error. In both cases, the bias was relatively small.

2. In all other conditions, the bias was substantial.

3. In CAT with item review, there is inward bias; that is, easy items are overestimated, and difficult items are underestimated.

4. If only simulees with $\theta > 0$ review the items, the bias in the easy items is reduced.

5. Choosing the complete test to be optimal at the true θ completely contaminates the calibration in the sense that all item parameters shrink to zero.

TABLE 3
Squared bias and standard errors for calibration β —20 items per test, 1,000 respondents

Item Selection Mode	β	Bias	S.E.	Mean
Random Selection	–2.0	0.01	0.32	–1.96
	–1.0	0.05	0.20	–0.94
	0.0	0.01	0.23	–0.01
	1.0	0.01	0.29	1.01
	2.0	0.08	0.29	2.08
CAT	–2.0	0.02	0.19	–2.00
	–1.0	0.03	0.26	–1.03
	0.0	0.01	0.08	–0.01
	1.0	0.00	0.21	0.99
	2.0	0.00	0.19	2.00
CAT with Item Review	–2.0	0.64	0.22	–1.33
	–1.0	0.34	0.29	–0.65
	0.0	0.01	0.07	–0.01
	1.0	0.28	0.22	0.71
	2.0	0.60	0.18	1.39
CAT with Item Review if $\theta > 0.0$	–2.0	0.07	0.21	–1.90
	–1.0	0.15	0.29	–0.84
	0.0	0.01	0.07	0.01
	1.0	0.20	0.19	0.79
	2.0	0.52	0.22	1.47
50% optimal at true θ	–2.0	0.43	0.23	–1.54
	–1.0	0.38	0.25	–0.61
	0.0	0.00	0.17	0.00
	1.0	0.33	0.22	0.66
	2.0	0.40	0.22	1.59
100% optimal at true θ	–2.0	1.92	0.39	–0.05
	–1.0	0.92	0.21	–0.07
	0.0	0.04	0.18	0.04
	1.0	0.93	0.22	0.06
	2.0	1.84	0.38	0.15
50% Initial Responses at $\hat{\theta}$ with s.d.($\hat{\theta}$) = 1.0	–2.0	0.21	0.20	–1.76
	–1.0	0.17	0.19	–0.82
	0.0	0.00	0.17	0.00
	1.0	0.08	0.21	0.91
	2.0	0.24	0.20	1.75
50% Initial Responses at $\hat{\theta}$ with s.d.($\hat{\theta}$) = 2.0	–2.0	0.21	0.19	–1.79
	–1.0	0.17	0.18	–0.95
	0.0	0.00	0.18	0.00
	1.0	0.08	0.21	0.93
	2.0	0.24	0.23	1.77

References

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

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

Eggen, T. J. H. M., & Verhelst, N. D. (1992). *Item calibration in incomplete testing designs* (Measurement and Research Department Reports, 92-3). Arnhem, The Netherlands: CITO.

Fischer, G. H., & Scheiblechner, H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12*, 23–51.

Glas, C. A. W. (1988). The Rasch Model and multi-stage testing. *Journal of Educational Statistics, 13*, 45–52.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27*(4), 887–903.

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

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177–195.

Mislevy, R. J., & Chang, H. H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65*, 149–156.

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

Mislevy, R. J., & Wu, P. K. (1996). *Missing responses and Bayesian IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report 96-30-0NR). Princeton, NJ: Educational Testing Service.

Neyman, J., & Scott, E. L. (1948). Consistent estimates from partially consistent observations. *Econometrica, 16*, 1–32.

Rigdon, S. B., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48*, 567–574.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23–48.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International, Inc.