RR-88-48-ONR

INFERRING EXAMINEE ABILITY WHEN SOME ITEM RESPONSES ARE MISSING

Robert J. Mislevy

and

Pao-Kuei Wu
University of California at Berkeley
Educational Testing Service

This research was sponsored in part by the
Cognitive Science Program
Cognitive and Neural Sciences Division
Office of Naval Research, under
Contract No. N00014-85-K-0683

Contract Authority Identification No.
NR 150-539

Robert J. Mislevy, Principal Investigator

Educational Testing Service
Princeton, New Jersey

September 1988

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No 0704-0188

1a REPORT SECURITY CLASSIFICATION: Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S): RR-88-48-ONR	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION: Educational Testing Service; 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION: Cognitive Science Program, Office of Naval Research (Code 1142PT), 800 North Quincy Street
6c ADDRESS (City, State, and ZIP Code): Princeton, NJ 08541	7b ADDRESS (City, State, and ZIP Code): Arlington, VA 22217-5000
8a NAME OF FUNDING/SPONSORING ORGANIZATION; 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-85-K-0683
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
61153N	RR04204	RR04204-01	NR 150-539

11 TITLE (Include Security Classification): Inferring Examinee Ability When Some Item Responses Are Missing (Unclassified)

12 PERSONAL AUTHOR(S): Robert J. Mislevy and Pao-Kuei Wu

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Technical	FROM _____ TO _____	September 1988	68

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Adaptive testing; Omitted responses; Item response theory; Targeted testing; Missing data
05	10		

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The basic equations of item response theory (IRT) provide a foundation for inferring examinees' abilities and items' operating characteristics from observed responses. In practice, though, examinees will usually not have provided a response to every available item--for reasons that may or may not have been intended by the test administrator, and that may or may not be related to examinee ability. The mechanisms that produce missingness must be taken into account if correct inferences are to be drawn. Using concepts introduced by Rubin (1976), we discuss the implications for ability and item parameter estimation that are entailed by alternate test forms, targeted testing, adaptive testing, time limits, and omitted responses.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL: Dr. Charles E. Davis	22b TELEPHONE (Include Area Code): 202-696-4046; 22c OFFICE SYMBOL: ONR 1142CS

DD Form 1473, JUN 86 — Previous editions are obsolete — S/N 0102-LF-014-6603

Inferring Examinee Ability When Some Item Responses Are Missing

Robert J. Mislevy

Educational Testing Service

Pao-Kuei Wu

University of California at Berkeley

Educational Testing Service

September 1988

This paper results from work carried out during the second author's summer internship at ETS. The first author was supported by Contract No. N00014-85-K-0683, project designation NR 150-539, from the Cognitive Science Program, Cognitive and Neural Sciences Division, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government. We are grateful to Murray Aitkin and Kentaro Yamamoto for their comments and suggestions.

Inferring Examinee Ability When Some Item Responses Are Missing

Abstract

The basic equations of item response theory (IRT) provide a foundation for inferring examinees' abilities and items' operating characteristics from observed responses. In practice, though, examinees will usually not have provided a response to every available item--for reasons that may or may not have been intended by the test administrator, and that may or may not be related to examinee ability. The mechanisms that produce missingness must be taken into account if correct inferences are to be drawn. Using concepts introduced by Rubin (1976), we discuss the implications for ability and item parameter estimation that are entailed by alternate test forms, targeted testing, adaptive testing, time limits, and omitted responses.

Key words: Adaptive testing; Item response theory; Missing data; Omitted responses; Targeted testing

Introduction

The capability to measure different examinees with different test items is an oft-cited advantage of item response theory (IRT). This option implies a problem of inference in the presence of missing data, since an examinee may not have provided a response to every item in the complete item set. Five types of missingness are in fact encountered regularly in routine applications of IRT:

Case 1: Alternate test forms. Two or more tests with similar content but different items are often employed to minimize carry-over effects (as in test-retest designs), reduce fatigue and practice effects (by splitting a test into shorter subtests), or avoid cheating behavior. A examinee is typically administered one form selected at random.

Case 2: Targeted testing. Two or more tests with similar content, but pitched at different levels of difficulty, can be used to make testing more efficient when background information (such as grade or courses taken) is available for deciding which test to administer to each examinee.

Case 3: Adaptive testing. Testing can also be made more efficient and less time-consuming if each item presented to an examinee is selected on the basis of his responses up to that point, and possibly background information as well.

Case 4: Not-reached items. Under typical testing conditions, some examinees will not reach the last few items on a test because of the time limit.

Case 5: Omitted items. Even when an item has been presented to an examinee and he has time to reach it, he will sometimes choose not to respond.

When incomplete data of any of these types are encountered, the IRT model that presumably accounts for the responses that are observed, is embedded in a more encompassing model that determines which responses will be observed and which will be missing. This paper discusses the implications that missing responses hold for likelihood and Bayesian inferences about examinee ability parameters and item parameters, assuming an IRT model holds. When can the process that causes missingness be ignored? When it cannot be ignored, how can it be modeled? How can conventional IRT methods for missing responses be evaluated in this framework?

The following section extends IRT notation to handle missingness, using concepts and notation from Little and Rubin (1987) and Rubin (1976). Next, Rubin's (1976) conditions for when the missingness process can be ignored are reviewed. Each of the five types of missingness listed above are then discussed in some detail in the problem of inferring ability when item

parameters are known. This is followed by the extension to item parameter estimation. A final section summarizes our results.

Background and Notation

At the heart of IRT is the model for the response to item j, with its possibly vector-valued parameter β_j, from an examinee with ability θ. The Rasch model for dichotomous items, for example, posits

$$P(U_j = u_j|\theta,b_j) = \exp[u_j(\theta - b_j)]/[1 + \exp(\theta - b_j)] ,$$

where u_j=1 denotes a correct response and u_j=0 an incorrect one, and b_j is the difficulty parameter of item j. We assume IRT functions that are twice differentiable, and interpret $P(U_j = 1|\theta,\beta_j)$ as the proportion of correct responses we would expect to many items with $\beta = \beta_j$ from many examinees with that value of θ.

Under the usual assumption of local independence, the conditional probability of the response vector $U = (U_1,\ldots,U_n)$ for n items is obtained by the product rule:

$$P(U = u|\theta,\beta) = \prod_{j=1}^{n} P(U_j = u_j|\theta,\beta_j) . \qquad (1)$$

It is further assumed that if y denotes background information about an examinee such as age, GPA, or courses taken, then

$$P(U = u|\theta,\beta,y) = P(U = u|\theta,\beta) \ .$$

When there is no possibility of missing responses, (1) can be interpreted as a likelihood function, say $L(\theta|\underset{\sim}{u})$, once a particular value $\underset{\sim}{u}$ of U has been observed. Direct likelihood inferences are based solely on relative values of L at different values of θ. It might be said, for example, that the probability of $\underset{\sim}{u}$ is twice as high at θ' than at θ''. The maximum likelihood estimate (MLE), $\hat{\theta}$, is the value at which $\underset{\sim}{u}$ has the highest probability. Note that in direct likelihood inference, the MLE concerns only the data that were actually observed.

The role of the MLE in sampling distribution inferences concerns its distribution under repeated sampling of observations with a fixed "true" parameter value. If n is large, the sampling distribution of $\hat{\theta}$ as computed from repeated observations of U can be approximated by a normal distribution with mean θ and variance

$$\sigma^2 = -\left[\frac{\partial^2 \ell(\theta|\underset{\sim}{u})}{\partial\theta^2}\right]^{-1}$$

where $\ell(\theta|\underset{\sim}{u}) = \log L(\theta|\underset{\sim}{u})$. By considering the distribution of $\hat{\theta}$ over hypothetical draws from the sample space, sampling distribution inferences involve datasets that could have been observed, but were not.

Bayesian inferences are based on the posterior distribution for θ given $\underline{u}$, or

$$p(\theta|\underline{u}) = K\, L(\theta|\underline{u})\, p(\theta)\ , \qquad (2)$$

where K is a normalizing constant and $p(\theta)$ conveys knowledge about θ before a value of U is observed. The posterior mean and mode of θ are sometimes taken as point estimates in IRT. The posterior variance is approximated by σ^2 when n is large. (This is the variance of the posterior distribution for θ induced by the data actually observed, in contrast to the variance of an estimator over hypothetical repeated observations).

In many applications of IRT, an examinee provides responses to only a subset of the n items to which responses could have been observed. The data thus consist of (i) the identification of the subset of items to which responses are observed and (ii) the responses to those items. The first inferential problem we address is to estimate an individual examinee's θ from this extended observation, assuming that both the IRT model and the item parameters are known. To this end, we adapt notation from Little and Rubin (1987) and Rubin (1976) in defining the following terms:

- $\underline{U} = (U_1, \ldots, U_n)$ is the (hypothetical) random vector of responses to all items in the full item set.

- $M = (M_1,\ldots,M_n)$ is an associated "missing-data indicator," with each element taking values of 0 or 1. If $m_j = 1$, the value of U_j will be observed; if $m_j = 0$, the value of U_j will be missing.

- $V = (V_1,\ldots,V_n)$ conveys the data that are actually observed: $V_j = U_j$ if $m_j = 1$ but $V_j = *$ if $m_j = 0$.

An observed value of M, say $\underset{\sim}{m}$, effects a partition of U, u, V, and v according to which elements are observed and which are missing. That is, we may write $U = (U_{mis},U_{obs})$ to distinguish the missing and observed elements of U, respectively. Similarly, $u = (u_{mis},u_{obs})$, $V = (V_{mis},V_{obs})$, and $v = (v_{mis},v_{obs})$. As with $\underset{\sim}{u}$ and $\underset{\sim}{m}$, let v denote a realized value of V.

Example

An examinee is administered a two-item test. With each item scored right or wrong (1 or 0), there are $2^2 = 4$ possible patterns for U: (0,0), (0,1), (1,0), and (1,1). The second response may be missing, however. With 1 representing "observed" and 0 representing "missing," there are 4 conceivable patterns for M, of which (1,0) and (1,1) can be realized. If the examinee would have responded incorrectly to the first item and correctly to the second, but the response for the second item is missing, then $\underset{\sim}{u} = (0,1)$, $\underset{\sim}{m} = (1,0)$, and $\underset{\sim}{v} = (0,*)$. #

Inferences must of course be based on the data that are actually observed, namely realizations of $V = (U_{obs}, M)$. Modeling the hypothetical complete data vector (U,M)--even if there is no intention of observing a response to every item--is a convenient way to begin. It forces us to explicate our beliefs about the relationships among ability, item response, and missingness--exactly what is required for building a sensible model for V. Recalling that p(u,m) can be written as $p(m|u)\ p(u)$ or as $p(u|m)\ p(m)$, define the following densities:

- $f_\theta(u)$ is the density for all n responses. In this paper, $f_\theta(u)$ takes the form shown in (1), so by local independence we can write $f_\theta(u) = f_\theta(u')\ f_\theta(u'')$ for any ordering and partitioning of the items into (u',u'')--including (u_{mis}, u_{obs}).

- $g_\phi(m|u)$ is the probability that M takes the value $m = (m_1, \ldots, m_n)$ given that U takes the value $u = (u_1, \ldots, u_n)$, with ϕ being the (possibly vector-valued) parameter of the missingness process. It is possible for θ to be a component of ϕ, in which case the value of θ itself plays a role in determining whether a response will be observed. In these cases we shall sometimes write $g(m|u,\theta,\phi)$ to emphasize the dependence on θ.

- $h_\theta(u|m)$ is the probability that U takes the value u given that M takes the value m.

- $t_\phi(m)$ is the probability that M takes the value m. Again, θ may be a component of ϕ.

Example (continued)

Suppose that the missingness process in the two-item example initiated above can be described as follows: The second response is observed whenever the first response is correct; the second response will be observed with probability ϕ if the first response is incorrect. Then

$$g_\phi(m|u) = \begin{cases} 1 & \text{if } m=(1,0) \text{ and } u=(1,0) \text{ or } (1,1) \\ 1-\phi & \text{if } m=(1,0) \text{ and } u=(0,0) \text{ or } (0,1) \\ \phi & \text{if } m=(1,1) \text{ and } u=(0,0) \text{ or } (0,1) \\ 0 & \text{otherwise.} \end{cases} \quad \#$$

Whenever not all potential responses may be observed for any reason--even if they all do turn out to be observed--the data are v. To obtain the likelihood function, we start with the likelihood for the (hypothetical) complete data (U,M), then average over the missing responses u_{mis}:

$$L(\theta,\phi|\underset{\sim}{v}) = \delta[(\theta,\phi),\Omega_{\theta\phi}] \int f_\theta(u_{mis},\underset{\sim}{u}_{obs}) g_\phi(\underset{\sim}{m}|u_{mis},\underset{\sim}{u}_{obs}) du_{mis} ,$$

where δ takes the value 1 if a value (θ,ϕ) is in the parameter space $\Omega_{\theta\phi}$ and 0 if it isn't. This observed-data likelihood is a weighted average over all complete-data likelihoods that have the targeted responses to the observed items. The weights are proportional to the probabilities of these potential response patterns for the different values u_{mis}, given $\underset{\sim}{m}$ and $\underset{\sim}{u}_{obs}$. Using local independence, we can bring the probability for the observed responses outside the integral:

$$L(\theta,\phi|\underset{\sim}{v}) = \delta(\cdot,\cdot)\ f_\theta(\underset{\sim}{u}_{obs}) \int f_\theta(u_{mis}) g_\phi(\underset{\sim}{m}|u_{mis},\underset{\sim}{u}_{obs}) du_{mis} . \quad (3)$$

Equivalently, using the alternative expression for p(u,m),

$$L(\theta,\phi|\underset{\sim}{v}) = \delta(\cdot,\cdot)\ t_\phi(\underset{\sim}{m}) \int h_\theta(u_{mis},\underset{\sim}{u}_{obs}|\underset{\sim}{m})\ du_{mis} . \quad (4)$$

Appropriate likelihood inferences are based on relative values of $L(\theta,\phi|\underset{\sim}{v})$ at various values of (θ,ϕ), or at various values of θ after eliminating ϕ by conditioning or maximizing. Appropriate Bayesian inferences are based on the posterior distribution

$$p(\theta,\phi|\underset{\sim}{v}) \propto L(\theta,\phi|\underset{\sim}{v})\ p(\theta,\phi) , \quad (5)$$

where $p(\theta,\phi)$ conveys prior knowledge about θ and ϕ. Appropriate sampling distribution maximum likelihood inferences concern the

distribution of $(\hat{\theta},\hat{\phi})$ from (3) or (4), over hypothetical repeated observations of V for fixed (θ,ϕ).

In general, then, the correct likelihood function involves a nuisance parameter ϕ, and depends not on just the responses that were observed, through $f_\theta(\tilde{u}_{obs})$, but on the responses that were not observed, through $f_\theta(u_{mis})$ and $g_\phi(m|u_{mis},\tilde{u}_{obs})$.

Example (continued)

With IRT for binary variables, the integral over u_{mis} that appears in (3) is a summation over all possible response patterns with $u_{obs} = \tilde{u}_{obs}$. In our two-item example with the first response incorrect and the second response missing, the potential complete patterns u are (0,1) and (0,0). Thus,

$$L(\theta,\phi|V=(0,*)) = \delta[(\theta,\phi),\Omega_{\theta\phi}]\ f_\theta(U_1=0)$$

$$\times\ \{f_\theta(U_2=0)\ g_\phi[M=(1,0)|U=(0,0)]$$

$$+\ f_\theta(U_2=1)\ g_\phi[M=(1,0)|U=(0,1)]\}\ . \qquad (6)\ \#$$

Conditions for Ignorability

Ignoring the missingness process when drawing inferences about θ means that rather than using the correct likelihood

$L(\theta,\phi|\underset{\sim}{v})$, using a facsimile of (1) as applied to $\underset{\sim}{u}_{obs}$ alone:

$$L^*(\theta|\underset{\sim}{u}_{obs}) = \delta(\theta,\Omega_\theta)\ f_\theta(\underset{\sim}{u}_{obs})$$

$$= \delta(\theta,\Omega_\theta) \prod_{obs} P(U_j = \underset{\sim}{u}_j|\theta,\beta_j) \ . \qquad (7)$$

In particular, direct likelihood inferences about θ that ignore the missingness process simply compare values of L^* at various values of θ. Bayesian inferences that ignore the missingness process start with an analogue of (2), a psuedo-posterior distribution proportional to

$$L^*(\theta|\underset{\sim}{u}_{obs})\ p(\theta) \ . \qquad (8)$$

Sampling-distribution maximum likelihood inferences that ignore the missingness process consider the distribution of $\hat{\theta}$ from (7) over repeated samples of responses to the items for which $\underset{\sim}{m}_j$=1. This involves a different reference sample space--not the sample space of v values, driven by (θ,ϕ), but a sample space of $\underset{\sim}{u}_{obs}$ values for a fixed $\underset{\sim}{m}$, driven by θ. (This reasoning is used in survey sampling when the exact size of the sample is not known before it is obtained. Even though the sample size N is a random variable with its own distribution and parameters, standard errors for $\hat{\theta}$ are typically computed with respect to repeated draws with the observed sample size $\underset{\sim}{N}$, rather than with respect to repeated draws of (U_{obs},N).)

It is a pleasant state of affairs when ignoring the missingness process leads to the correct inferences about θ, since (7) and (8) don't require the specification of g, h, or t, and standard computing algorithms can be used. Depending on why the missing responses were missing, however, these procedures need not lead to the correct inferences. Rubin (1976, 1987) specifies conditions under which a missingness process can be ignored under sampling distribution, direct likelihood, and Bayesian inference. They involve the concepts **missing at random**, **missing completely at random**, and **distinctness of parameters**:

Definition 1: Missing responses are **missing at random** (MAR) if for each value of ϕ and for all fixed values m and u_{obs}, $g_{\phi}(m|u_{mis},u_{obs})$ takes the same value for all u_{mis}. (This definition of MAR applies to the missingness process in general, as in Rubin, 1987, rather than a specific value of the missingness variable, as in Rubin, 1976.)

Definition 2: Missing responses are **missing completely at random** (MCAR) if for each value of ϕ and for each fixed value m, $g_{\phi}(m|u)$ takes the same value for all u.

Definition 3: The parameter θ is **distinct** (D) from ϕ if their joint parameter space factors into a θ-space and a ϕ-space, and when prior distributions are specified for θ and ϕ, they are independent.

Taken together, MCAR and D imply that the values of both the observed and the missing responses are independent of the pattern of missingness. MAR and D together imply that the values of the missing responses are independent of the pattern of missingness, conditional on the values of the observed responses. MCAR implies MAR.

Example (continued)

In order to satisfy MCAR, it must be that for each value of ϕ and any value of m, $g_\phi(m|u)$ takes the same value for all u. In our two-item example, however, $g_\phi(M=(1,0)|U=(1,1))=1$ while $g_\phi(M=(1,0)|U=(0,0))=1-\phi$. Except in the trivial case that $\Omega_\phi=\{0\}$, MCAR is not satisfied.

In order to satisfy MAR, it must be that for each value of ϕ and any fixed values of m and u_{obs}, $g_\phi(m|u_{mis},u_{obs})$ takes the same value for all values of u_{mis}. This condition is satisfied trivially whenever m=(1,1), since there are no missing observations. It is also satisfied trivially in our example when m=(0,1) or m=(0,0), since these missingness patterns have probability zero for all u. The following equalities for m=(1,0) complete the verification of MAR:

$$g_\phi(M=(1,0)|U=(1,0)) = g_\phi(M=(1,0)|U=(1,1)) = 1$$

$$g_\phi(M=(1,0)|U=(0,0)) = g_\phi(M=(1,0)|U=(0,1)) = 1-\phi \ . \qquad \#$$

We are now in a position to summarize Rubin's conclusions regarding direct likelihood and Bayesian inference. First, a more easily verified sufficient condition:

- When making direct-likelihood or Bayesian inferences about θ, it is appropriate to ignore the process that causes missing data if missing data are missing at random and the parameter of the missing data process is "distinct" from θ. (Rubin, 1976, p. 581)

When MAR is satisfied, g does not depend on u_{mis} and can be brought out of the integral in (4), which then simply integrates to one. If D is satisfied as well, the only dependence of $L(\theta,\phi|\tilde{v})$ on θ is through $L^*(\theta|\tilde{u}_{obs})$ $\left[= f_\theta(u_{obs})\right]$.

Under weaker conditions for ignorability, the integral need not drop out as it does under MAR, but its value does not depend on θ. Necessary and sufficient conditions are as follows:

- Suppose $L^*(\theta|\tilde{u}_{obs}) > 0$ for all $\theta \in \Omega_\theta$. All likelihood ratios for θ ignoring the process that causes missing data are correct for all $\phi \in \Omega_\phi$, if and only if (a) $\Omega_{\theta\phi} = \Omega_\theta \times \Omega_\phi$ and (b) for each $\phi \in \Omega_\phi$,

$$E_{u_{mis}}[g_\phi(\tilde{m}|u_{mis},\tilde{u}_{obs})|\tilde{m},\tilde{u}_{obs},\theta,\phi] \qquad (9)$$

takes the same positive value for all θ. (From Rubin, 1976, Theorem 7.2.)

o The posterior distribution of θ ignoring the process that causes missing data equals the correct posterior distribution of θ if and only if

$$E_{\phi,u_{mis}}[g_{\phi}(\underset{\sim}{m}|u_{mis},\underset{\sim}{u}_{obs})|\underset{\sim}{m},\underset{\sim}{u}_{obs},\theta] ,$$

or

$$\int\int g_{\phi}(\underset{\sim}{m}|u_{mis},\underset{\sim}{u}_{obs})\ p(u_{mis}|\theta)\ p(\phi|\theta)\ d\phi du_{mis} ,$$

takes a constant positive value. (From Rubin, 1976, Theorem 8.2.)

Example (continued)

Equation 6 gives the complete-data likelihood for the observed data $\underset{\sim}{v} = (0,*)$, namely $L(\theta,\phi|V = (0,*))$. When does the psuedo-likelihood $L^{*}(\theta|U_1=0)$ yield the same direct likelihood inferences about θ? For this to happen, it must first hold that the θ and ϕ sample spaces are distinct; it cannot be, for instance, that the observed pattern of missingness could occur for some values of θ but not for others. Second, the following term that appears in (6) must be constant for all values of θ:

$$f_{\theta}(U_2=0)\ g_{\phi}[M=(1,0)|U=(0,0)] + f_{\theta}(U_2=1)\ g_{\phi}[M=(1,0)|U=(0,1)] .$$

MAR would mean that $g_\phi[M=(1,0)|U=u]$ is constant for all u, in which case the expression simplifies to

$$[f_\theta(U_2=0) + f_\theta(U_2=1)]\ g_\phi[M=(1,0)]\ ;$$

then, since the sum in brackets is one, simply to the constant value $g_\phi[M=(1,0)]$. When this happens, the sufficient condition for ignorability is satisfied. But even if $g_\phi[M=(1,0)|U=u]$ is not constant over u, the entire expression can be constant for θ if the variations in f and g over θ cancel each other out. For example, it could be that $\phi = \theta$ and, for $u_2 = 0,1$,

$$g_\phi[M=(1,0)|U=(0,u_2)] = [f_\theta(U_2=u_2)]^{-1}\ .$$

As we shall see in the case of intentional omissions, such constraints are not generally plausible in the context of IRT. #

When ignorability under direct likelihood inference holds for a given missingness process--as occurs when MAR is satisfied--the correct value of θ is identified as the MLE. The usual sampling-distribution interpretation of $\hat{\theta}$ and σ^2 may or may not be justified. (Recall that if the sampling interpretation is to be justified at all, it will be with respect to repeated response sampling with m fixed at $\tilde{m}$. The variance of $\hat{\theta}$, for example, may be quite different in this frame of reference from its variance under repeated samples of (u_{obs},m).)

First, a sufficient condition:

- When making sampling-distribution inferences about statistics T(v), it is appropriate to ignore the process that causes missing data if missing data are missing completely at random and the parameter of the missing data process is "distinct" from θ. (Little and Rubin, 1987, p. 14)

Under these conditions, $p(U=u,M=m|\theta,\phi) = f_\theta(u)\, g_\phi(m)$ with θ and ϕ distinct, and $\underset{\sim}{v}$ may be thought of as the outcome of a two-stage experiment: ϕ determines $\underset{\sim}{m}$ in the first stage and θ determines $\underset{\sim}{u}_{obs}$ in the second. An experimenter looking at the results of the second stage has the same information about θ as an experimenter who has performed that latter experiment only with the value $\underset{\sim}{m}$ predetermined.

A necessary and sufficient condition for ignorability for sampling inferences about a generic statistic T(v), based on Rubin's (1976) Theorem 6.2, is as follows:

- The sampling distribution of T(v) under f_θ calculated by ignoring the process that causes missing data equals the correct conditional sampling distribution of T(v) given m under under f_θ and g_ϕ if and only if for each fixed value of m,

$$E_{u_{mis}}[g_{\phi}(m|u)|m,u_{obs},\theta,\phi] - E_{u}[g_{\phi}(m|u)|m,\theta,\phi] > 0 \ .$$

Equivalently, the probability of each missingness pattern must not depend on the values of the responses that are observed; for each fixed m and any u'_{obs} and u''_{obs}, it must be true that

$$Pr(M=m|U_{obs}=u'_{obs},\theta,\phi) = Pr(M=m|U_{obs}=u''_{obs},\theta,\phi) \ .$$

This condition is implied by Rubin's (1976) slightly stronger "observed at random." Unless it holds, $(\underset{\sim}{u}_{obs},m)$ does not admit to a decomposition into a sequence of independent experiments because the value of U_{obs} plays a role in determining M, and the conditional frame of reference is not appropriate.

Inferences about Examinee Ability

The following sections address in turn the common types of missingness in IRT that were listed in the introduction, in the problem of drawing inferences about θ when β is known. In each case, we consider whether the conditions for ignoring missingness are plausible, and, when they are not, discuss how the missingness process might be modeled so that inferences can be drawn.

Case 1: Alternate test forms

By "alternate test forms," we mean tests whose items all fit the same IRT model, and which provide information sufficiently

similar that the test administrator is indifferent as to which form any particular examinee is presented. The form an examinee receives will depend a random process such as a coin flip or a form-spiralling scheme. The common practice in IRT applications with alternate test forms is to base inferences about θ on $L^*(\theta|\tilde{u}_{obs})$.

The use of K alternate test forms implies that only K missingness patterns, $m_1, \ldots, m_k, \ldots m_K$, can occur, where all the elements of m_k are zero except those that correspond to the items that appear in Form k. Denote their respective probabilities by $\phi_k = P(M=m_k)$. Assuming the IRT model means that $f_\theta(U)$ is as given in (1); that is, we assume that item responses would be governed by θ alone, regardless of which items would be presented. Even though the items of only one form will actually be presented, it is possible to express our assumptions about the connection between the (hypothetical) values of complete response pattern and the probability of the missingness pattern as follows:

$$g_\phi(m|u) = \begin{cases} \phi_k & \text{for all u if } m=m_k\text{; i.e., Form k} \\ 0 & \text{otherwise.} \end{cases}$$

Since the values of g do not depend on u, MCAR, and therefore MAR, are satisfied. Verifying D for likelihood inference requires that all values of θ are possible with all possible values of ϕ; they are. Verifying D for Bayesian inference requires that prior beliefs about θ and ϕ be independent; this is eminently reasonable

as well. Having satisfied the sufficient conditions MCAR and D, we conclude that the missingness caused by the random administration of alternate test forms is ignorable under direct likelihood and Bayesian inference, and under sampling distribution interpretations of $\hat{\theta}$. Common practice is therefore justified.

Case 2: Targeted Testing

Targeted testing also involves multiple test forms, but ones in which the distributions of item difficulty differ from form to form. Exploiting the fact that estimates of θ are more precise when an examinee is administered items with difficulties near θ, targeted testing uses background information y about an examinee to select a test form that will probably be more informative about him than other forms. For example, an easy form and a hard form might be constructed from a set of n items calibrated together under the same IRT model, then the easy form could be given to first graders and the hard form to second graders.

As in Case 1, the existence of K forms implies that only K patterns of M, namely $m_1, \ldots, m_K$, can be realized. The parameter of the missingness process has values $\phi_k(y)$, which indicate the probability that an examinee with background variable y will be administered Form k. For at least one k and two values y' and y'', $\phi_k(y') \neq \phi_k(y'')$; this happens when $p(\theta|y') \neq p(\theta|y'')$, and the difficulty of Form k is better suited to the typical examinee with one value of y than the other. If we denote the easy and hard forms in the two-form example mentioned above as

1 and 2, and let y denote grade level (1 or 2), then $g_{\phi}(M=m_k|U=u,y) = \phi_k(y)$ for all u, with

$$\phi_k(y) = P(M=m_k|y) = \begin{cases} 1 \text{ if } y = 1 \text{ and } k = 1 \\ 0 \text{ if } y = 1 \text{ and } k = 2 \\ 0 \text{ if } y = 2 \text{ and } k = 1 \\ 1 \text{ if } y = 2 \text{ and } k = 2 \end{cases}.$$

Because g does not depend on u, MCAR is satisfied. Assuming that all values of θ are <u>possible</u> at all values of y--even if they are more likely for some values of y than others--distinctness as required for direct likelihood inference is also satisfied. The values of maximum likelihood estimates of θ from $L^*(\theta|\tilde{u}_{obs})$ are therefore the correct values under targeted testing. This is all that matters for direct likelihood interpretation of the MLE. Sampling-distribution interpretations are also appropriate, with respect to repeated administrations of the form that actually was administered.

Distinctness as required for Bayesian inference is <u>not</u> satisfied. Prior beliefs about θ and ϕ are associated through y, so $p(\theta,\phi) \neq p(\theta)\,p(\phi)$. Intuitively, knowing which form an examinee was administered under targeted testing is a source of information about θ since form selection depends on prior knowledge about θ through y. This knowledge must be taken into account in Bayesian inference. It is true, however, that $p(\theta,\phi|y) = p(\theta|y)\,p(\phi|y)$. Bayes-distinctness is satisfied <u>conditional on</u>

$\underline{y}$, and the missingness process can be ignored conditional on $\underline{y}$. Thus, the correct Bayesian inferences under targeted testing are obtained with $L^*(\theta|\tilde{u}_{obs})\ p(\theta|y)$, but generally not with $L^*(\theta|\tilde{u}_{obs})\ p(\theta)$.

Case 3: Adaptive Testing

As mentioned for Case 2, IRT measurement can be made more efficient by presenting an examinee with items that are informative in the neighborhood of his θ. Adaptive testing uses information from an examinee's preceding responses, and possibly from from his background variables y as well, to select each next item to administer. As responses accumulate, more is known about θ and successive item selections are more accurately targeted.

The datum observed in adaptive testing is a sequence of nobs ($\leq$n) ordered pairs, $S = ((I_1,U_{obs(1)}),\ldots,(I_{nobs},U_{obs(nobs)}))$, where I_k identifies the k'th item administered and $U_{obs(k)}$ is the response to that item. Define the partial response sequence S_k as the first k ordered pairs in S, with the null sequence s_0 representing the status as the test begins. Augment the set of n items with the fictitious Item 0, the selection of which corresponds to a decision to terminate testing. It can be written as the nobs+1'st item in the test, although no response is associated with it.

A test administrator defines an adaptive test design by specifying for all items j, all realizable partial response sequences s_k, and all values of y, the probabilities $\phi(j,s_k,y)$

that item j will be selected as the k+1'st test item, after observing the partial response sequence s_k from an examinee with Y=y. The dependence of item selection probabilities upon y allows for a hard item to be the first one presented to a high school graduate, say, but an easy one to be first for a nongraduate. Item selection probabilities in designs that do not use y can be written simply as $\phi(j,s_k)$. We begin by considering designs of this latter type.

One example of an adaptive testing design is Bayesian minimum variance item selection (Owen, 1975). In its pure form, the item that minimizes the expected posterior variance of θ, using the current posterior distribution $p(\theta|s_k)$, is chosen as the $k+1^{st}$ item with probability one. To reduce the exposure of more informative items, positive probabilities may instead be assigned to several fairly informative items. Typically, testing continues until either a desired level of precision is reached, or a predetermined number of items has been administered.

A second example of an adaptive testing design is the two-item example employed earlier in this paper. Its definition of g corresponds to administering the first item to all examinees, and with probability ϕ, the second item to some of the examinees who answered the first item incorrectly.

In adaptive testing, the probability of observing s from an examinee with ability θ can be built up sequentially. The probability of selection for the first item is $\phi(i_1,s_0)$. The probability of response $u_{obs(1)}$ to item i_1 is given by the IRT

model as $f_\theta(u_{obs(1)})$--by conditional independence, a value that does not depend on the fact that this item happened to have been presented first. The conditional probability of selection for the second item <u>given s_1</u> is $\phi(i_2,s_1)$--a value independent of θ. The probability of the corresponding response is $f_\theta(u_{obs(2)})$, a value independent of the identification of, and the response to, the first item. Continuing in this manner until it is determined to stop testing, with probability $\phi(0,s)$, we obtain

$$P(s|\theta) = \prod_{k=1}^{nobs+1} \phi(i_k,s_{k-1}) \prod_{k=1}^{nobs} f_\theta(u_{obs(k)}) .$$

The likelihood function induced by the observation of $\tilde{s}$ is thus

$$L(\theta|\tilde{s}) = \prod_{k=1}^{nobs+1} \phi(\tilde{i}_k,\tilde{s}_{k-1}) \; L^*(\theta|\tilde{u}_{obs}) . \qquad (11)$$

Observe that...

1. $\tilde{s}$ conveys the value of $\tilde{m}$: $\tilde{m}_j = 1$ if $i_k = j$ for some k, $1 \leq k \leq nobs$; otherwise, $\tilde{m}_j = 0$.

2. $\tilde{s}$ conveys the value of $\tilde{u}_{obs}$, namely the responses to the items administered during the course of the test.

3. L factors into two products, the first of which depends on ϕ, $\underset{\sim}{u}_{obs}$, and $\underset{\sim}{m}$, and the second of which--namely $L^*(\theta|\underset{\sim}{u}_{obs})$ --depends on θ and $\underset{\sim}{u}_{obs}$.

4. If s' and s'' imply the same m and u_{obs}, then $L(\theta|s') \propto L(\theta|s'')$.

Points 3 and 4 justify the use of $L^*(\theta|\underset{\sim}{u}_{obs})$ for direct likelihood inference. It may be instructive nonetheless to verify the satisfaction of MAR. Now P(M=m,U=u), or the probability of the hypothetical complete observation (m,u), is the probability of observing a response sequence s that yields the targeted m and u_{obs}, times the probability that the unobserved responses u_{mis} also take the targeted values. Defining $T = \{s: M=m \cap U_{obs}=u_{obs}\}$ as the set of response sequences that present the targeted items and have the targeted responses to them, we have

$$P(M=m,U=u) = P(s\epsilon T)\ P(U_{mis}=u_{mis})$$

$$= \left\{ \sum_{T} \prod_{k}^{nobs+1} \phi(i_k,s_{k-1}) \prod_{obs} f_\theta(u_{obs(k)}) \right\} \left\{ \prod_{mis} f_\theta(u_{mis(k)}) \right\}$$

$$= \left\{ \sum_{T} \prod_{k}^{nobs+1} \phi(i_k,s_{k-1}) \right\} \prod_{j=1}^{n} f_\theta(u_j)$$

$$= \left(\sum_{T} \prod_{k}^{nobs+1} \phi(i_k, s_{k-1}) \right) P(U=u) \ .$$

But then

$$g_{\phi}(m|u) = P(M=m, U=u)/P(U=u)$$

$$= \sum_{T} \prod_{k}^{nobs+1} \phi(i_k, s_{k-1}) \ ,$$

a value that does not depend on u_{mis}, as required for MAR. This argument also holds when ϕ depends on y. (QED)

MAR and distinct parameter spaces are sufficient for ignorability of the adaptive-testing missingness mechanism under direct likelihood inference. Ignorability holds under Bayesian inference if, in addition, the prior distributions for θ and ϕ are independent. As with targeted testing, this latter condition fails if for some y' and y'' for which $p(\theta|y') \neq p(\theta|y'')$, there exist j and s such that $\phi(j,s,y') \neq \phi(j,s,y'')$. When this is so, Bayesian inference demands the use of $p(\theta|y)$ rather than $p(\theta)$ in conjunction with $L^{*}(\theta|\tilde{u}_{obs})$.

Even though ignorability under direct likelihood inference means that L^{*} yields the correct maximizing value from a given observation, sampling-distribution interpretation of the MLE $\hat{\theta}$ is

not justified in general. To see this, recall that the necessary condition for sampling-distribution ignorability requires that for each fixed m and any u'_{obs} and u''_{obs},

$$Pr(M=m|U_{obs}=u'_{obs},\theta,\phi) = Pr(M=m|U_{obs}=u''_{obs},\theta,\phi) \ .$$

This would require that the probability of any given missingness pattern be the same no matter what values the responses took. But since by definition adaptive tests produce missingness patterns as a function of the response values that are observed, only a degenerate adaptive testing scheme could satisfy this condition.

Concluding that the item selection mechanism is not ignorable for sampling distribution inference means that the correct sampling distribution for $\hat{\theta}$ must be calculated with respect to repeated administrations of the entire adaptive test. While general theory does not relate its variance in this frame of reference to the second derivative of L^*, the latter may be a reasonable approximation of the former under particular adaptive test designs. Whether this is so must be determined individually for each adaptive test design, analytically in simple cases but by simulation in more realistic cases.

Case 4: Not-Reached Items

IRT is intended for "power" tests, or those in which an examinee's chances of responding correctly would not differ appreciably if the time limit were more generous. Time limits are

typically chosen to allow most examinees to respond to all items, but a few examinees won't have time to answer all of them. This section concerns the items that an examinee does not reach. It assumes the examinee has not interacted with the item--e.g., he has not seen what the items at the end of the test ask, and decided to work instead on the ones he has seen at the beginning of the test.

It is common practice to identify not-reached items by working from the end of an examinee's response string toward the beginning, taking unanswered items as not-reached until an answer is encountered. Unanswered items preceding this last answered item are taken as intentional omissions, and will be considered in the next section. Concentrating on nonresponse due to not-reached only, and limiting our attention to examinees who have reached at least the first the item, we must address n patterns of missingness: for $\ell = 0,\ldots,n-1$, let m_ℓ denote the string of $n-\ell$ 1's followed by ℓ 0's. That is, m_ℓ is the missingness pattern of an examinee that has not reached the last ℓ items.

Checking ignorability. We continue to assume that a common IRT model holds for the responses of items reached, u_{obs}, and not-reached, u_{mis}. This assumption is crucial for applying IRT models to data with not-reached items, and two ways of checking it will be discussed at the end of the section. When it does hold, the missingness process is characterized by the examinee speed parameter $\phi = (\phi_0,\ldots,\phi_{n-1})$ of a multinomial variable, where ϕ_ℓ is the probability that missingness pattern m_ℓ will be observed--

i.e., that the last ℓ items will not be reached. Under this formulation, the probability of the complete observation (m,u) is obtained as

$$p(m,u|\theta,\phi) = p(m|\phi) \prod_{j}^{n} p(u_j|\theta) \ ;$$

MCAR (and therefore MAR) holds. The probability of v is

$$p(v|\theta,\phi) = p(m|\phi)\ p(u_{obs}|\theta) \int p(u_{mis}|\theta)\ du_{mis}$$

$$= p(m|\phi)\ p(u_{obs}|\theta) \ . \qquad (12)$$

If, in addition to MCAR, all values of θ are possible at all values of ϕ--even if some are more likely than others--the not-reached missingness process is ignorable with respect to direct likelihood inference. That is, direct likelihood inference about θ in the presence of not-reached items can be based on $L^*(\theta|u_{obs})$. Sampling-distribution inferences about θ from $\hat{\theta}$ are also appropriate. They pertain to repeated sampling of responses to the items that were reached, and enjoy the asymptotic sampling properties of MLEs if the number of items reached is large.

For ignorability to hold under Bayesian inference, it is necessary in addition to MAR that $p(\theta,\phi) = p(\theta)\ p(\phi)$; that is, that "speed" and "ability" are independent. Empirical evidence suggests that this is not generally true. Van den Wollenberg (1979), for example, reports significant positive correlations between percent-correct scores on the first eleven items (which

were reached by all examinees) and the total number of items reached, in four of six intelligence tests in the ISI battery (Snijders, Souren, and Welten, 1963). Bayesian inference about θ would take this relationship into account by using the correct posterior distribution

$$p(\theta|\tilde{v}) \propto \int L(\theta,\phi|\tilde{v})\, p(\theta,\phi)\, d\phi$$

$$= \int L^{*}(\theta|\tilde{u}_{obs})\, L(\phi|\tilde{m}) p(\phi|\theta)\, d\theta\, p(\theta)$$

$$= L^{*}(\theta|\tilde{u}_{obs})\, p(\theta) \int L(\phi|\tilde{m})\, p(\phi|\theta)\, d\theta$$

$$\propto L^{*}(\theta|\tilde{u}_{obs})\, p(\theta|\tilde{m}) \ .$$

Checking the IRT model. Verifying MCAR for not-reached items required assuming that the responses that would have been observed, had those items been reached, follow the same IRT model as those that were reached. We now describe two ways of checking this assumption, one using only the response data $\tilde{v}$ that are normally observed, the other requiring the researcher to discover not-reached responses in a supplemental data-gathering effort.

A necessary condition for an IRT model to hold in the presence of not-reached items, is that the same IRT model hold for reached items among examinees who have reached different numbers of items. Let u_{obs} be the observed responses to items that are reached in a sample of N examinees, and let $nobs_i$ be the number of items examinee i reaches. Let β_j be the parameter(s) of item j. The marginal probability of u_{obs} under the hypothesis that item parameters are invariant over groups of examinees with different missingness patterns is

$$P_A(u_{obs}) = \prod_i \int \prod_{j=1}^{nobs_i} p(u_{ij}|\theta,\beta_j)\, p(\theta|m_i)\, d\theta \ . \qquad (13)$$

Viewing (13) as a likelihood function and maximizing with respect to $\beta_1,\ldots,\beta_n$ yields the value L_A.

An alternative hypothesis is that item parameters vary over not-reached groups. We can estimate n-j+1 different item parameters for item j, where $\beta_{j\ell}$ applies to those examinees who have reached $n-\ell$ items. For example, Item 3 will have parameters for groups who reached n, n-1, ..., 4, and 3 items. The marginal probability of u_{obs} under this hypothesis is

$$P_B(u_{obs}) = \prod_i \int \prod_{j=1}^{nobs_i} p(u_{ij}|\theta,\beta_{j,nobs_i})\, p(\theta|m_i)\, d\theta \ ,$$

which leads to the maximizing likelihood value L_B. In large samples, $-2 \log(L_A/L_B)$ is approximately chi-square with $n(n - 1)/2$ degrees of freedom when the null hypothesis is true.

Van den Wollenberg (1979) provides empirical evidence that the "item parameter invariance" with respect to not-reached groups is often, but not always, tenable. Applying his own goodness-of-fit indices rather than the likelihood ratio suggested above, he verified this type of invariance in five of the six ISI tests.

A second way of studying the IRT assumption in the presence of not-reached items begins by finding out what the responses to the not-reached items would have been. This can be accomplished with paper-and-pencil tests by allowing a sample of examinees to continue beyond the usual time limit until they have answered every item, but using a different colored pencil after the usual limit. Of the total of n items, then, examinee i will have responded to the first $nobs_i$ under the normal time limits and the remaining $nmis_i = n - nobs_i$ thereafter. Under the null hypothesis of an invariant IRT model across reached and not-reached items, the marginal probability of the completed response matrix $u = (u_{mis}, u_{obs})$ under the null hypothesis is

$$p_C(u_{obs}) = \prod_i \int \prod_j p(u_{ij}|\theta,\beta_j)\, p(\theta|m_i)\, d\theta \ .$$

Alternatively, we can fit an IRT model that allows both the item parameters and examinee parameters to differ before and after the time limit. Each item except the first can have two parameters, $\beta_{j,mis}$ and $\beta_{j,obs}$, whenever some examinees answered before the time limit and some answered it after; each examinee can two abilities, θ_{mis} and θ_{obs}. The resulting marginal probability is

$$P_D(u_{obs}) = \prod_i \int\int \left[\prod_{j=1}^{nobs_i} p(u_{ij}|\theta_{obs},\beta_{j,obs_i}) \right]$$

$$\times \left[\prod_{j}^{nmis_i} p(u_{ij}|\theta_{mis},\beta_{j,mis_i}) \right] p(\theta_{obs},\theta_{mis})\, d\theta_{obs}\, d\theta_{mis} \quad .$$

In large samples, $-2 \log(L_C/L_D)$ is approximately chi-square under the null hypothesis, with degrees of freedom equal to the number of items with two parameters appearing in (14), plus the number of additional parameters estimated for the examinee parameter distribution $p(\theta_{obs},\theta_{mis})$ over those required for $p(\theta)$.

Case 5: Omitted Responses

A missing response is an intentional omission when the examinee is administered the item, has time to appraise its content, and decides for his own reasons not to make a response.

After showing that such omissions can't generally be considered ignorable, we discuss a number of ways to deal with them.

Omitting behavior. Test scores T(v) are assigned to patterns of rights, wrongs, and omits for the purposes of comparing or selecting examinees. Assuming that a correct response to an item gives a higher value of T than an incorrect response and that an examinee wants to obtain a high score, he will make responses he believes are correct. How he will respond to an item about which he is unsure depends at least partly on how the test will be scored (Sabers and Feldt, 1968).

Formula scores, for example, take the form

$$T(v) = R(v) - X\,W(v) ,$$

where R(v) and W(v) are counts of right and wrong responses and X is a constant selected by the test administrator. Setting X = 0 gives number-right scores; X = 1 gives right-minus-wrong scores; for multiple choice items with A alternatives, X = 1/(A-1) gives the familiar "corrected-for-guessing" scores. The examinee maximizes his expected score by answering items for which he thinks his chances of being correct are at least X/(1 + X). In particular, he should answer every item under number-right scoring, and those for which he thinks his chances are at least c = 1/A under corrected-for-guessing scoring. Some examinees either do not use this strategy, or make inaccurate assessments of their

chances. Analyzing responses to items that examinees originally omitted under right-minus-wrong scoring, Sherriffs and Boomer (1954) did find about half, or X/(1 - X), the omitted responses would have been correct among examinees who scored low on a risk-aversion scale, but nearly two-thirds would have been correct among examinees with high risk-aversion scores.

The examinee's perceived probabilities of correct response must be distinguished from the probabilities of the IRT model. IRT gives the proportion of correct response to Item j from examinees with ability θ, but each of these examinees may have a different idea of his own chances. They may differ in the accuracy of their estimates and their confidence about them, and their perceived probabilities need not average to the IRT probability. Observing whether an examinee omits an item merely tells us something about what he thinks u_j would be.

Are omits ignorable? To see that the assumptions needed for ignorability are not generally plausible, we examinee the case in which n = 1; i.e., a single item. MAR simplifies to $g_\phi(m|U=0) = g_\phi(m|U=1)$, meaning that an omit is just as likely if the response would have been correct as if it would have been incorrect. But since examinees tend to answer items they feel are correct, MAR implies the unappealing assumption that their perceptions of correctness are independent of actual correctness.

MAR (along with D) is merely sufficient for ignorability, however, and ignorability can hold when MAR does not. For a single item, the necessary condition for ignorability under

likelihood inference given in (9) requires that for each value of ϕ, the expression

$$g_{\phi}(M=0|U=0)\ f_{\theta}(U=0) + g_{\phi}(M=0|U=1)\ f_{\theta}(U=1) \qquad (15)$$

take the same value for all θ. This is the just the probability that the item will be omitted. That its value remains constant as θ increases without limit flouts intuition, since we'd expect examinees whose high abilities virtually assure a correct response to be aware of their high chances, and respond rather than omit. This conjecture is borne out in empirical studies such as Stocking, Eignor, and Cook (1988) that show markedly lower rates of omission from examinees with high (corrected-for-guessing) scores than from examinees with low scores.

Since ignorability is not satisfied for direct likelihood inference, L^{*} does not generally yield the correct MLE, and sampling distribution inferences based on the resulting value are inappropriate. The requirements for ignorability under Bayesian inference are the same as those for direct likelihood inference except that they must apply when averaged over ϕ rather than for each particular value; ignorability is thus implausible there too.

Lord (1974) argues against ignoring omits under maximum likelihood scoring, saying that the examinee who knew we planned to use the MLE from $L^{*}(\theta|\bar{u}_{obs})$ as a score would omit all items except those for which he was certain his response would be

correct. This plausible argument also presumes a relationship between actual and examinee-perceived item correctness.

Filling in the blanks. Lord (1974) suggested that omits on multiple-choice items under guessing-corrected scoring can be handled with standard IRT estimation routines if they are treated as as fractionally correct, with value c. He assumed "rational" omitting behavior: examinees omit items only if their chances of responding correctly would have been c, so that $h_{\phi}(U_j=1|M_j=0) = c$ for all items and all θ. Omitting decisions are also assumed to be independent from one item to the next, given θ and ϕ. In a natural extension of conditional independence of item responses given θ, we assume "extended local independence," or conditional independence of item responses and missingness given θ and ϕ:

$$P(U=u,M=m|\theta,\phi) = \prod_j P(U_j=u_j,M_j=m_j|\theta,\phi) \quad .$$

Under these assumptions, the complete-data likelihood takes the following form:

$$L(\theta,\phi|\tilde{u},\tilde{m}) = \prod_j^n p(u_j,m_j|\theta,\phi)$$

$$= \prod_j^n f_{\theta}(u_j)g(m_j|u_j,\theta,\phi)$$

$$= \prod_{j}^{n} P_j(\theta)^{\tilde{u}_j} Q_j(\theta)^{1-\tilde{u}_j} \prod_{j}^{n} g(\tilde{m}_j|\tilde{u}_j,\theta,\phi)$$

[where $P_j(\theta) = f_\theta(U_j=1)$ and $Q_j(\theta) = 1 - P_j(\theta)$]

$$= L_u(\theta|\tilde{u})\ L_{m|u}(\theta,\phi|(\tilde{m}|\tilde{u}))\ .$$

The complete-data likelihood thus factors into two terms, with L_u being the IRT-based probability of item responses and $L_{m|u}$ the probability of the missingness pattern given the response pattern. Both depend on θ. Were u and m both fully observed, the usual MLE based on L_u would be a conditional MLE, foregoing the additional information conveyed by $L_{m|u}$ but avoiding the nuisance parameter ϕ. One would proceed by finding the maximizing value of the log likelihood

$$\ell_u = \sum_{j}^{n} \tilde{u}_j \log P_j(\theta) + (1 - \tilde{u}_j) \log Q_j(\theta)\ .$$

The same conditional-estimation strategy can be applied to the observed data $\tilde{v} = (\tilde{u}_{obs},\tilde{m})$, by maximizing the conditional expectation of L_u, or $E[L_u(\theta|u_{mis},\tilde{u}_{obs})|(\tilde{u}_{obs},\tilde{m})]$. Finding the maximizing θ for $E(L_u)$ by Dempster, Laird, and Rubin's (1977) EM

algorithm requires finding the maximizing θ for the expectation of ℓ_u, given $\underset{\sim}{u}_{obs}$, $\underset{\sim}{m}$, and a provisional estimate θ^0; that is, of

$$F(\theta|\theta^0) = \int \ell_u(\theta|u_{mis},\underset{\sim}{u}_{obs})\, p(u_{mis}|\underset{\sim}{u}_{obs},\underset{\sim}{m},\theta = \theta^0)\, du_{mis}$$

$$= \sum_{obs} \underset{\sim}{u}_j \log P_j(\theta) + (1 - \underset{\sim}{u}_j) \log Q_j(\theta)$$

$$+ \int[\sum_{mis} u_j \log P_j(\theta) + (1-u_j)Q_j(\theta)]\, p(u_{mis}|\underset{\sim}{u}_{obs},\underset{\sim}{m},\theta=\theta^0)\, du_{mis}. \tag{16}$$

But under Lord's assumptions,

$$p(u_{mis}|\underset{\sim}{u}_{obs},\underset{\sim}{m},\theta = \theta^0) = \prod_{mis} h_\theta(u_j|m_j = 0)$$

$$= \prod_{mis} c^{u_j}(1 - c)^{1-u_j},$$

a value that doesn't depend on ϕ or θ^0. Substituting this expression into the second term of (16), the integral simplifies to

$$F(\theta|\theta^0) = \sum_{obs} \tilde{u}_j \log P_j(\theta) + (1 - \tilde{u}_j) \log Q_j(\theta)$$

$$+ \sum_{mis} c \log P_j(\theta) + (1 - c) \log Q_j(\theta) . \quad (17)$$

This is equivalent to the log of Lord's Equation 4, the psuedo-likelihood obtained by using $u_j = c$ in the complete-data likelihood whenever $m_j = 0$. Equation (17) does not depend on θ^0, so the EM algorithm comprises only a single cycle. Maxima of (17) are maxima of $E(L_u)$. A global maximum is assured if the complete-data probability belongs to the exponential family, as is the case with the Rasch model.

Lord (1974) points out that the criterion function obtained by replacing omits with fractionally-correct responses is not the likelihood function induced by the observation. We have shown, however, that the resulting estimate of θ maximizes what might be called a "marginal conditional" likelihood function, allowing one to apply standard results from the theory of maximum likelihood estimation, such as consistency--in this context, as the number of items not omitted increases.

The foregoing analysis yields insight into other treatments of omits that impute values for u_{mis}. Supplying random responses that are correct with probability c provides a crude numerical evaluation of (16), leading to a maximizing value whose

expectation is the value obtained when the integration is carried out in closed form as in (17). This practice is justified by the same assumptions as Lord's (1974) approach, but sacrifices accuracy for convenience. Supplying incorrect responses for omits leads to a "marginal conditional" MLE for θ under the assumption that responses to omitted items would surely have been incorrect. This may be reasonable for open-ended items, but it is not plausible for multiple-choice items for which even the least able examinees have nontrivial probabilities of success. In these cases, supplying incorrect responses for omits would bias estimates of θ downward.

Lord addressed "rational" omitting behavior, in that the expectation of correctness for an omitted response is always c, the value associated with the optimal omitting strategy. As we have noted however, studies of responses to items originally omitted show that not all examinees behave in this manner. The tendency to omit when probabilities of success may be higher than c can be associated with personality characteristics, demographic variables, and level of ability. This approach biases estimates of θ downward for risk-aversive examinees. We now discuss how such dependencies can be taken into account, although it is by no means certain that this should be done; to do so effectively adjusts scores upward or downward in accordance with examinee background characteristics, which may be objectionable on the grounds of fairness. Assuming rational omitting behavior in scoring rules, and making the rules and optimal strategies as

clear as possible to examinees, may be preferable when test scores are used to make sensitive placement or selection decisions.

Modeling empirical rather than ideal omitting behavior requires a study like that of Sherriffs and Boomer's (1954), where examinees are first administered a test under standard conditions, then later asked what their responses to the items they omitted would have been. From these data it is possible to calculate proportions of omits that would have been correct as a function of the items and examinee characteristics--possibly including θ. If θ is not included, empirical estimates $h(U_j=1|M_j=0,y)$ are employed in place of c in (16). This takes into account possible differences in rates of omitted correct response from one item to the next--some higher than c, some lower--or among examinees with different demographic or measured psychological characteristics. If θ is included, then estimates of θ employing $h_\theta(U_j=1|M_j=0,y)$ must be calculated iteratively. The values $h_\theta(U_j=1|M_j=0,y)\Big|_{\theta=\theta^0}$ replace c in (17) for each missing response, and an improved estimate θ^1 is obtained via maximum likelihood. This must then be used to produce an improved estimate of the expectation of each missing response, $h_\theta(U_j=1|M_j=0,y)\Big|_{\theta=\theta^1}$. From these and $\bar{u}_{obs}$, yet another estimate θ^2 will be obtained. The process is repeated until convergence occurs. The original estimation of item parameters and of the functions $h_\theta(U_j=1|M_j=0,y)$ requires similar modifications to standard item parameter estimation algorithms. Additional

parameters for $h_\theta(U_j=1|M_j=0,y)$ can be estimated jointly with standard item parameters; a plausible implementation would make the logits of h's associated with each item linear in θ.

Lord's (1983) model for omits. While Lord's (1974) treatment of omits as fractionally correct yields reasonable and statistically defensible (conditional) MLEs of θ when rational omitting behavior is assumed, the full likelihood induced by the data was neither presented nor exploited. To accomplish this requires an explication of the missingness process, in the form of a model for the joint probability distribution of U and M. Such a model was proposed by Lord (1983).

Lord's (1983) model for omits maintains the context of guessing-corrected scoring of multiple-choice items with $A(=1/c)$ alternatives, but offers additional structure for the response process. It is first assumed that an examinee either feels a preference for one of the alternatives or is totally undecided among them. The proportion of examinees with ability θ feeling no preference on Item j is $R_j(\theta)$. If a preference is felt, a response is made; of the responses made by examinees with ability θ who feel a preference, the proportion correct is $P^*_j(\theta)$. If no preference is felt, the examinee will either omit the item with probability ω or respond completely at random. Responses and omitting decisions are again assumed to be independent from one item to the next, conditional on θ and ω.

These assumptions imply that the missingness parameter ϕ is (θ,ω), and lead to the following forms for $t_\phi(m_j)$ and $h_\theta(u_j|m_j)$:

$$t_\phi(m_j) = \omega\, R_j(\theta)^{1-m_j}\,[1 - \omega R_j(\theta)]^{m_j}$$

and

$$h_\theta(u_j|m_j) = \begin{cases} c^{u_j}\,(1-c)^{1-u_j} & \text{if } m_j = 0 \\ P_j^{**}(\theta)^{u_j}\,[1 - P_j^{**}(\theta)]^{1-u_j} & \text{if } m_j = 1, \end{cases}$$

where $P_j^{**}(\theta) = P(U_j=1|\theta,m_j=1)$, the conditional probability that an observed response will be correct, is the sum of the probabilities of responding correctly when a preference is felt and guessing correctly when a preference is not felt:

$$P_j^{**}(\theta) = [1 - R_j(\theta)]\,P_j^{*}(\theta) + c(1-\omega)R_j(\theta)\ .$$

The joint likelihood for θ and ω induced by $\underset{\sim}{v}$ (i.e., $\underset{\sim}{u}_{obs}$ and $\underset{\sim}{m}$) is thus

$$L(\theta,\omega|\underset{\sim}{v}) = t_{\phi}(\underset{\sim}{m}) \int \prod_{j}^{n} h_{\theta}(u_j|\underset{\sim}{m}_j)\, du_{mis}$$

$$= t_{\phi}(\underset{\sim}{m}) \prod_{obs} h_{\theta}(\underset{\sim}{u}_j|M_j = 1) \int \prod_{mis} h_{\theta}(u_j|M_j = 0) du_{mis}$$

$$= \prod_{j}^{n} [\omega R_j(\theta)]^{1-\underset{\sim}{m}_j} [1 - \omega R_j(\theta)]^{\underset{\sim}{m}_j} \prod_{obs} P_j^{**\underset{\sim}{u}_j}(\theta) Q_j^{**}(\theta)^{1-\underset{\sim}{u}_j} . \tag{18}$$

Assuming the functions P^* and R are known, (18) provides a basis for full-information inference about θ. Under maximum likelihood, the joint maximum for (θ,ω) may be found by standard numerical methods, and a large-sample variance estimate can be based on the inverse of the second derivative of the log of (18) with respect to θ and ω. Under Bayesian inference, the posterior for θ and ω is obtained by multiplying (18) by $p(\theta,\omega)$ and normalizing; from this point, one may examine characteristics of the joint posterior for θ and ω, or integrate ω out to obtain the marginal posterior for θ.

Lord suggests that this model might be implemented by specifying functional forms for P^* and R, e.g., the 3-parameter logistic IRT function for P^* and the 2-parameter logistic with a negative slope for R. The underlying model for the correctness of item responses, observed or not, can be written as a function of P^* and R as follows:

$$P(U_j=1|\theta,\omega)$$

$$= P(U_j=1|\theta,\omega,M_j=1)\ P(M_j=1|\theta,\omega) + P(U_j=1|\theta,\omega,M_j=0)\ P(M_j=0|\theta,\omega)$$

$$= \{c(1-\omega)R_j(\theta) + [1-R_j(\theta)]P_j^*(\theta)\}\ [1-\omega R_j(\theta)] + c\omega R_j(\theta)\ .$$

Note that this probability depends on ω as well as θ; thus, the underlying model for item responses U_j is not a standard IRT model depending on θ alone and exhibiting local independence. This would be true only if for each value of θ, all examinees with that θ had the same value of ω. A special case of this requirement is for all examinees at all values of θ to have the same value of ω. Lord points out that if this were so with $\omega = 0$--i.e., no propensity toward omitting, even when no preference is felt--the resulting IRT model would be

$$P_j(\theta) = P_j^*(\theta)\ [1-R_j(\theta)] + c\ R_j(\theta)\ .$$

In a manner described by Samejima (1979), a response curve of this form need not be monotonically increasing over the range of θ. High-ability examinees would tend to feel preferences and respond correctly; moderate-ability examinees might tend to feel a preference for a clever distractor and answer incorrectly at a rate lower than c; very low ability examinees would feel no preference at all, and answer correctly at a rate equal to c.

Nominal category models. IRT models for multiple-category items have been proposed by Bock (1972), Samejima (1979), Sympson (1983), Thissen and Steinberg (1986), and others. These models have sometimes been used for data with intentional omissions, with an omit treated as one more possible response to a multiple-choice item. Lord (1983) expresses reservations about this practice,

> "...since it treats probability of omitting as dependent only on the examinee's ability, whereas it actually depends on a dimension of temperament. It seems likely that local unidimensional independence may not hold." (p. 477)

The following analysis makes Lord's concerns more explicit.

The features of the approach regarding omission are retained when all overt incorrect responses are collapsed into a single category. Recalling that the values 0, 1, and * of v stand for observed wrong, observed right, and omit, we obtain the multiple-category model probabilities f^*_θ as follows:

$$f^*_\theta(V_j = 0) = P(U_j = 0, M_j = 1|\theta)$$

$$= \int f_\theta(U_j = 0)\, g(M_j = 1|U_j = 0, \theta, \phi)\, p(\phi|\theta)\, d\phi \ ;$$

$$f^*_\theta(V_j = 1) = P(U_j = 1, M_j = 1|\theta)$$

$$= \int f_\theta(U_j = 1)\, g(M_j = 1|U_j = 1, \theta, \phi)\, p(\phi|\theta)\, d\phi \ ;$$

and

$$f^{*}_{\theta}(V_j = *) = P(U_j = 0, M_j = 0|\theta) + P(U_j = 1, M_j = 0|\theta)$$

$$- \int f_{\theta}(U_j = 1)\, g(M_j = 0|U_j = 1, \theta, \phi)\, p(\phi|\theta)\, d\phi$$

$$+ \int f_{\theta}(U_j = 0)\, g(M_j = 0|U_j = 0, \theta, \phi)\, p(\phi|\theta)\, d\phi \ .$$

Under the assumption of "extended local independence,"

$$p(U = u, M = m|\theta, \phi) = \prod_j p(U_j = u_j, M_j = m_j|\theta, \phi) \ .$$

This implies

$$p(V = v|\theta, \phi) = \prod_j p(V_j = v_j|\theta, \phi) \ . \qquad (19)$$

Using (19),

$$p(V = v|\theta) = \int p(V = v|\theta, \phi)\, p(\phi|\theta)\, d\phi$$

$$= \int \prod_j p(V_j = v_j|\theta, \phi)\, p(\phi|\theta)\, d\phi \ . \qquad (20)$$

But for local independence to hold in the usual sense for the multiple-category model, it would be necessary that

$$p(V = v|\theta) = \prod_j \int p(V_j = v_j|\theta,\phi)\ p(\phi|\theta)\ d\theta\ ,$$

and this does not generally follow from (20) as the order of multiplication and integration has been interchanged. It does follow if for each θ value, ϕ takes the same value for all examinees with that θ value; that is, the variables of the omitting process may vary from one value of θ to another, but not among examinees with the same value of θ. Lord's objection, then, may be stated as a desire to allow for different propensities for omitting to occur within a given level of ability.

A second reservation that might be offered for this approach stems from the fact that probabilities for v given θ are averages over ϕ. Even if (i) $f_\theta(u)$ is an IRT model satisfying local independence and (ii) $f_\theta(u)\ g_\phi(m|u)$ satisfies extended local independence, the multiple-category response curves $f^*_\theta(v)$ will vary from one group of examinees to the next unless the conditional distributions $p(\phi|\theta)$ are invariant over groups.

How to model omits, if you must. Standard IRT concerns examinees' tendencies to make correct responses when omits cannot occur. When they can occur, the differences among examinees' tendencies to omit responses can be cast as a nuisance variable in the classic sense. It is often easier to deal with such extraneous influences at the time the data are collected than to model them after the fact. In aptitude and ability testing, we should inform examinees as clearly as possible how their

performance will be evaluated, and persuade them as convincingly as possible to use the omitting strategy that maximizes their expected score. To the degree we succeed, variation in ϕ is reduced and examinees' data differ mainly because of differences in θ. If too the proportion of omits is low, imputing fractionally-correct or even random responses at the level c yields inferences that are plausible, readily-calculable, and robust with respect to alternative models for omitting.

For the sake of completeness, however, we now outline an approach using a full model for response and omission. The model exhibits local independence for elements of U given θ, and extended local independence for elements of (U,M) given (θ,ϕ). Its implementation requires either that g is assumed to be known or that an experiment with the same items and similar examinees has revealed the values of item responses that were originally omitted. We assume here that the experiment has been carried out, and a complete data matrix $(\underset{\sim}{u},\underset{\sim}{m})$ is available for a sample of examinees from a population of interest.

An IRT model $f_\theta(U = u,\theta,\beta)$ is assumed for item responses. The missingness process is modeled in terms of functions $g(M_j = m_j|U_j = u_j,\theta,\phi,\eta_j)$ for each item, where η_j are now additional item parameters for the frequency with which the item is omitted. For example, we could estimate from the completed data set item-omitting parameters $\eta_j = (d_{j0},e_{j0},d_{j1},e_{j1})$ that give the logit regression of m_j on θ when $u_j = 0$ and when $u_j = 1$; that is,

$$\text{logit } P(M_j = 0 | U_j = 0, \theta, \eta_j) = d_{j0}\,\theta + e_{j0}$$

$$\text{logit } P(M_j = 0 | U_j = 1, \theta, \eta_j) = d_{j1}\,\theta + e_{j1} \ ,$$

where logit $P = \log(P/(1 - P))$. The examinee omitting parameter ϕ could then be a tendency to omit more or less than average, so that

$$\text{logit } P(M_j = 0 | U_j = 0, \theta, \phi, \eta_j) = d_{j0}\,\theta + e_{j0} + \phi$$

$$\text{logit } P(M_j = 0 | U_j = 1, \theta, \phi, \eta_j) = d_{j1}\,\theta + e_{j1} + \phi \ .$$

The complete-data likelihood function for the item parameters is

$$L(\beta, \eta | \underset{\sim}{u}, \underset{\sim}{m}) =$$

$$\prod_{i}^{N} \iint f_{\theta}(\underset{\sim}{u}_i | \theta, \beta)\ h(\underset{\sim}{m}_i | \underset{\sim}{u}_i, \theta, \phi, \eta)\ p(\theta, \phi)\ d\theta\ d\phi \ . \qquad (21)$$

Equation (21) provides a basis for estimating (β, η) from the experimental data, either directly via maximum likelihood or, after multiplication by a prior distribution, via Bayesian methods. Maximum likelihood yields point estimates $(\hat{\beta}, \hat{\eta})$; Bayesian methods yield the posterior distribution $p(\beta, \eta | \underset{\sim}{u}, \underset{\sim}{m})$. Estimating the examinee parameter distribution $p(\theta, \phi)$ at the same

time might also be desirable, say by positing a functional form and estimating its parameters.

The results of this calibration can then be used to estimate the θ value of new examinee i, from whom only $\tilde{v}_i$ is observed. Under Bayesian inference, the relevant posterior distribution is

$$p(\theta|\tilde{v}_i,\tilde{u},\tilde{m}) \propto$$

$$\iiint L(\tilde{v}_i|\theta,\phi,\beta,\eta)\ p(\beta,\eta|\tilde{u},\tilde{m})\ p(\theta,\phi)\ d\beta\ d\eta\ d\phi\ .$$

Under maximum likelihood inference, the maximizing value $(\hat{\theta},\hat{\phi})$ of $L(\theta,\phi|\tilde{v}_i,\beta=\hat{\beta},\eta=\hat{\eta})$ might be sought.

Inferences about Item Parameters

When not all item responses are observed, the (marginal) likelihood function for item parameters β induced by the data matrix $\tilde{v} = (\tilde{v}_1,\ldots,\tilde{v}_N)$ from a sample of N examinees is

$$L(\beta|\tilde{v})$$

$$= \prod_i^N \iiint p(u_{mis,i}, u_{obs,i}, m_i|\theta,\phi,\beta)\ du_{mis,i}\ p(\theta,\phi)\ d\theta\ d\phi$$

$$= \prod_i^N \iint p(\tilde{v}_i|\theta,\phi,\beta)\ p(\theta,\phi)\ d\theta\ d\phi\ , \qquad (22)$$

where $u_{obs,i}$ and $u_{mis,i}$ are the observed and missing portions of the response vector of examinee i. The psuedo-likelihood obtained by ignoring the missingness process is

$$L^{*}(\beta|u_{obs}) = \prod_{i}^{N} \int p(\tilde{u}_{obs,i}|\theta,\beta)\ p(\theta)\ d\theta \ . \qquad (23)$$

Ignoring the missingness process when making direct likelihood inferences about β means comparing the values of (23) rather than (22) at different values of β. Equation (22) differs from (23) by integrating over $p(\tilde{u}_{obs,i}|\theta,\beta)$ with respect to θ for each examinee, rather than over $p(\tilde{v}_{i}|\theta,\phi,\beta)$ with respect to θ and ϕ. The resulting integrals are proportional with respect to β if and only if for all values of β, the conditions for ignorability for θ given β are satisfied for Bayesian inference. Therefore,

- Ignorability under direct (marginal) likelihood inference about β holds if (i) ignorability under Bayesian inference holds for each θ conditional on β, and (ii) β, θ, and ϕ are distinct in the sense required for direct likelihood inference.

If β is a priori independent of all θ and ϕ, the correct Bayesian posterior for β in the presence of missing data is the product of (22) and the prior $p(\beta)$. Ignoring the missingness process when making Bayesian inferences about β means using instead the product of (23) and $p(\beta)$. The preceding result indicates when ignoring the missingness gives the correct likelihood. To obtain the correct posterior, then, we have:

- Ignorability holds under Bayesian inference about β if (i) ignorability holds under Bayesian inference for each θ conditional on β, (ii) β, θ, and ϕ are distinct in the sense required for Bayesian inference.

If ignorability holds under direct likelihood inference, L^* yields the correct value for $\hat{\beta}$. The necessary condition for sampling distribution inferences based on $\hat{\beta}$ requires that the probability of the missingness pattern--in this context, the distribution of counts of individual missingness patterns--not depend on the values of observed responses. This condition is implied by MCAR. When it and direct-likelihood ignorability hold, conditional sampling distribution inferences are appropriate. They pertain to sampling of item responses to the observed items from repeated subsamples of examinees with each observed missingness pattern, with subsample sizes fixed at the observed counts of those missingness patterns.

Case 1. We have seen that for alternate test forms, ignorability holds for Bayesian inferences about θ given β. Random assignment of test forms ensures MCAR. Therefore the responses to items on forms that are not administered can be ignored under direct likelihood inference about β, and under Bayesian inferences as well as long as prior distributions are independent. Sampling-distribution inferences can based on the MLE $\hat{\beta}$ with the understanding that they pertain to repeated sampling of examinees for each form in the sample sizes that were actually observed.

Case 2. In targeted testing, items believed to be easier are administered to examinees that are expected to have lower abilities, and items believed to be harder are administered to examinees expected to have higher abilities. Bayesian inferences about θ given β can ignore missingness only after conditioning on y, the collateral examinee variable used in test-form assignment. Correct inferences about β under direct likelihood inference thus require that $p(\theta)$ in (23) be replaced by $p(\theta|y_i)$, as well as that all values of β are possible (if not always likely) on all test forms. Bayesian inferences must additionally take into account the prior beliefs that led to the differential assignment of items to forms. Let $z = (z_1,\ldots,z_n)$ represent the collateral information about items used to make these assignments (e.g., pretest item difficulties or item content). Appropriate Bayesian inferences about β that account for the missingness process may be drawn from

$$p(\beta|\tilde{u}_{obs},y,z) \propto L^*(\beta|\tilde{u}_{obs},y)\ p(\beta|z).$$

As in Case 1, the distribution of counts of missingness patterns does not depend on the values of observed item responses, so conditional sampling-distribution inferences about β from $L^*(\beta|\tilde{u}_{obs},y)$ are appropriate. They pertain to repeated administrations of the observed counts of administered forms at each value of y, to samples of examinees with those y values.

Case 3. Conclusions similar to those of Case 2 hold for direct likelihood and Bayesian inference in adaptive testing. Ignorability holds for Bayesian inference about θ given β (again conditioning on y if collateral examinee variables are used in item selection), so direct likelihood inferences about β from L^* are justified (Verhelst and Veldhuijzen, 1987)--though not always satisfactory. The reason is that $p(\beta|z)$ is often very strong in practice; indeed, β is usually treated as known for the purpose of item selection. In this case, z_j could be the mean $\bar{\beta}_j$ and covariance matrix of an assumed normal distribution for β_j, and item selection would be based on $\bar{\beta}_j$. The data collected for a given item when it is administered adaptively tend to be from examinees in a relatively narrow band of ability. For binary items with more than one parameter, the number of examinees required for stable estimates may well exceed the number that can be tested in practice. Bayesian inference is preferable under these circumstances. Provisional item parameter estimates based on $p(\beta|z)$ may be used to administer items, then adaptively-acquired item responses can be used to produce an updated distribution

$$p(\beta|z,\tilde{u}_{obs}) \propto L^*(\tilde{u}_{obs}|\beta)\ p(\beta|z) \ .$$

Because the missingness process is ignorable under direct likelihood inference, the usual MLE $\hat{\beta}$ obtained by maximizing L^*

gives the correct point estimate for MLE-based sampling-distribution inferences about β. As in the section on estimating θ from adaptive tests, however, the necessary condition for ignorability under sampling-distribution inference is <u>not</u> satisfied--the probabilities of missingness patterns depends on the values of observed responses. MLE properties based on L^* need not apply to item parameter estimates obtained from adaptive test data. It may be that for some adaptive test designs, the usual variance estimates with m fixed at $\bar{m}$ are good approximations to the variances that would be obtained under repeated sampling of the entire adaptive test for N examinees, but this must be determined case by case.

<u>Case 4</u>. Recall that when some items at the end of a test are not reached, MAR holds for inferences about θ given β but Bayesian ignorability does not hold unless speed and ability are independent. Missingness due to time limitations, therefore, is not generally ignorable under any type of inference about β. Assuming there are no restrictions on the parameter space, drawing likelihood inferences about β requires one to replace $p(\theta)$ in (22) with $p(\theta|\bar{m}_i)$, where

$$p(\theta|\bar{m}_i) \propto p(\theta) \int p(M - \bar{m}_i|\phi)\ p(\phi|\theta)\ d\phi\ .$$

If, in addition, the test has been assembled to start with easy items and become harder, the prior information about items [say

$p(\beta|z) = \Pi\ p(\beta_j|z_j)]$ will be related to the chances that examinees will not reach particular items. This information must also be used in Bayesian inference about β. The appropriate posterior is

$$p(\beta|\bar{v},z) \propto p(\beta|z)\ \Pi \int f_\theta(\bar{u}_{obs,i}|\beta)\ p(\theta|\bar{m}_i)\ d\theta\ .$$

Case 5. The topic of inference about item parameters when examinees omit some items intentionally has already been broached in the discussion about estimating θ. Bayesian ignorability for θ given β does not generally hold, so missingness mechanisms must be specified and inferences about β must start with on the full likelihood (23). A number of approaches were discussed there, including imputing responses for omits, using a multiple-category IRT model, and fitting Lord's (1983) model for responses and omits. The approach that is most easily incorporated into standard IRT algorithms is to treat intentional omits as fractionally correct (Lord, 1974). Assuming that examinees omit only in accordance with the strategy that maximizes their expected score, this approach gives "marginal-conditional" MLEs that maximize

$$E[L_u(\beta|u_{mis},\bar{u}_{obs})|\bar{u}_{obs},\bar{m}] =$$

$$\prod_i \int \prod_{obs_i} P_j(\theta)^{u_{ij}} Q_j(\theta)^{1-u_{ij}} \prod_{mis_i} P_j(\theta)^{c} Q_j(\theta)^{1-c} \, p(\theta) \, d\theta \, . \tag{24}$$

Equation (24) is "conditional" in that it accounts for the influence of θ upon item responses given the pattern of missingness, but does not capture the role of θ in determining that pattern. It is "marginal" in that it is the expectation over U_{mis} given u_{obs} and m of the conditional likelihood $L_u(\beta|u)$ that would be maximized if all responses had been observed.

Summary

In practical applications of item response theory (IRT), there are several reasons that item responses may not be observed from all examinees to all test items. Ignoring the missingness process under direct likelihood inference means using a psuedo-likelihood that includes terms for only the responses that were observed, without regard for the processes by which they came to be observed. The resulting inferences are appropriate if the psuedo-likelihood is proportional to the correct likelihood that does account for the missingness process. In this case the correct point estimate of an MLE is obtained. Sampling-distribution inferences from the MLE are appropriate only if the missingness pattern does not depend on the values of the observed data. When this condition holds, sampling-distribution inferences can be drawn with regard to repeated samples of responses under

the observed pattern of missingness. The missingness process is ignorable with respect to Bayesian inference if the correct Bayesian posterior is proportional to the product of the psuedo-likelihood and an appropriate prior distribution.

For fives common types of missingness in IRT, we used Rubin's (1976) theorems to determine whether ignorability holds under direct likelihood and Bayesian inference about examinee parameters θ when item parameters β are known. In those cases in which the correct value of the MLE is obtained under direct likelihood inference, we asked whether sampling distribution inferences based on the MLE were appropriate. We then considered the analogous questions for inferences about β when the examinee parameters are eliminated by marginalization. Our findings are summarized below. Tables 1 and 2 highlight the results on ignorability.

Case 1: Alternate test forms. When an examinee is assigned one of several alternative test forms by a random process such as a coin flip or a spiralling scheme, the process that renders missing the responses to items on the forms not presented is ignorable for all three types of inference, both for estimating β and for estimating θ when β is known.

Case 2: Targeted testing. When collateral variables such as educational or demographic status are used to assign an examinee one of several tests that differ in their measurement properties, the resulting missingness on forms not given is ignorable under direct likelihood inference for θ given β, but not under Bayesian

inference unless the prior information about examinees that led to differential assignments is conditioned on. This information must be taken into account for both likelihood and Bayesian inferences about β; for Bayesian inference, prior information about β used to select items must additionally be taken into account. Sampling distribution inferences may be based on MLEs for β and for θ given β, conditional on the observed patterns of test form administration within values of the examinee variables used for targeting.

Case 3: Adaptive testing. The same conclusions for direct likelihood and Bayesian inference follow in adaptive testing, where assignment proceeds item by item in accordance with the values of responses to preceding items. Ignorability under direct likelihood inference means that the correct points are identified as MLEs of θ given β and of β, but the usual MLE properties under sampling-distribution inference need not hold because the probabilities of missingness patterns depend on the values of observed responses.

Case 4: Not-reached items. When some examinees do not interact with the last items on a nearly nonspeeded test, the not-reached process is ignorable with respect to direct likelihood inference about θ given β, and the MLE supports sampling distribution inferences that pertain to repeated administrations of the items that were actually reached. This missingness process is not ignorable under Bayesian inference unless speed and ability are

independent. And only then can direct (marginal) likelihood inferences about β ignore the missingness. Bayesian inferences about β further require that prior knowledge about item parameters be employed if it played a role in determining which items would not be reached, as when items are ordered from easy to hard.

Case 5: Omitted items. When examinees are presented items, have a chance to appraise their content, and decide for their own reasons not to respond, the missingness is not ignorable. Inferences must be drawn from a full model for the joint distribution of missingness and item response.

Not surprisingly, modeling this nonignorable nonresponse is difficult. Neither of the two most ambitious approaches proposed to date, namely Lord's (1983) model for omits and the use of multiple-category IRT models, handles the issue of local independence in a fully satisfactory manner. Under the assumption that examinees are perfect judges of their chances of responding correctly, and omit only if it is in accordance with the strategy that maximizes their expected score, Lord's (1974) treatment of omits as fractionally correct can be justified as providing the expectation of a conditional term in the full likelihood. This procedure is readily incorporated into standard complete-data IRT algorithms and avoids having to specify the full likelihood, but sacrifices information about examinee and item parameters conveyed by the observed pattern of missingness.

References

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

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

Little, R.J.A., and 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. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. Psychometrika, 48, 477-482.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

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

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Sabers, D.L., and Feldt, L.S. (1968). An empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. Journal of Educational Measurement, 5, 251-258.

Samejima, F. (1979). A new family of models for the multiple-choice item. Research Report 79-4. Knoxville, TN: Department of Psychology, University of Tennessee.

Sherriffs, A.C., and Boomer, D.S. (1954). Who is penalized by the penalty for guessing? Journal of Educational Psychology, 45, 81-90.

Snijders, J.T., Souren, C.J.M.H., and Welten, V.J. (1963). ISI-publikaties 1. Algemene Handleiding. Groningen: Wolters.

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

Sympson, J.B. (1983). A new IRT model for calibrating multiple choice items. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.

Thissen, D., and Steinberg, L. (1984). A response model for multiple-choice items. Psychometrika, 49, 501-519.

van den Wollenberg, A.L. (1979). The Rasch model and time-limit tests. Doctoral dissertation, University of Nijmegen.

Verhelst, N., and Veldhuijzen, N. (1987). Estimating the correlation matrix from binary variables in incomplete designs. Arnhem: CITO.

Table 1

Ignorability Results for Estimating θ Given β

Type of Missingness	Type of Inference		
	Direct Likelihood	Bayesian	Sampling Distribution*
Alternate Forms	Yes	Yes	Yes
Targeted Forms	Yes	Yes, given examinee variables	Yes
Adaptive Testing	Yes	Yes, given examinee variables if they are used	No
Not-Reached	Yes	No, unless speed and ability are independent	Yes
Intentional Omissions	No	No	No

* Conditional on the observed pattern of missingness.

Table 2

Ignorability Results for Estimating β After Marginalizing over θ

Type of Missingness	Type of Inference: Direct Likelihood	Bayesian	Sampling Distribution*
Alternate Forms	Yes	Yes	Yes
Targeted Forms	Yes, given examinee variables	Yes, given examinee and item variables	Yes, given examinee variables
Adaptive Testing	Yes, given examinee variables if they are used	Yes, given item variables and examinee variables if they are used	No
Not-Reached	No, unless speed and ability are independent	No, unless speed and ability are independent	No, unless speed and ability are independent
Intentional Omissions	No	No	No

* Conditional on the observed pattern of missingness.