September 1988
doi:10.1002/j.2330-8516.1988.tb00304.x

Inferring Examinee Ability When Some Item Responses Are Missing

Robert J. Mislevy[1] and Pao-Kuei Wu[1,2]
[1]Educational Testing Service
[2]University of California at Berkeley

The basic equations of item response theory (IRT) provide a foundation for inferring examinee's abilities and items' operating characteristics from observed responses. In practice, though, examinees will usually not have provided a reponse 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 limites, and omitted reponses.

Keywords: 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-ver 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 from selected at random.

Case 2: Targeted testing. Two or more tests with similar content, but pitched a t 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 items 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 by ignored? When it cannot be ig-

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 conmments and suggestions.

nored, 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

$$\mathbb{P}\left(U_j = u_j \mid \theta, b_j\right) = \frac{\exp\left\{u_j(\theta - b_j)\right\}}{1 + \exp\left\{\theta - b_j\right\}},$$

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 $\mathbb{P}\left(U_j = 1 \mid \theta, \beta_j\right)$ 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 $\boldsymbol{U} = (U_1, \ldots, U_n)^\mathsf{T}$ for n items is obtained by the product rule:

$$\mathbb{P}\left(\boldsymbol{U} = \boldsymbol{u} \mid \theta, \beta\right) = \prod_{j=1}^{n} \mathbb{P}\left(U_j = u_j \mid \theta, \beta_j\right). \qquad (1)$$

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

$$\mathbb{P}\left(\boldsymbol{U} = \boldsymbol{u} \mid \theta, \beta, y\right) = \mathbb{P}\left(\boldsymbol{U} = \boldsymbol{u} \mid \theta, \beta\right).$$

When there is no possibility of missing responses, Equation (1) can be interpreted as a likelihood function, say $L\left(\theta \mid \widetilde{\boldsymbol{u}}\right)$, once a particular value $\widetilde{\boldsymbol{u}}$ of $\boldsymbol{U}$ 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 $\widetilde{\boldsymbol{u}}$ is twice as high at θ' than at θ''. The maximum likelihood estimate (MLE), $\widehat{\theta}$, is the value at which $\widetilde{\boldsymbol{u}}$ 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 $\widehat{\theta}$ as computed from repeated observations of $\boldsymbol{U}$ can be approximated by a normal distribution with mean θ and variance

$$\sigma^2 = -\left[\frac{\partial \ell\left(\theta \mid \widetilde{\boldsymbol{u}}\right)}{\partial\theta\partial\theta^\mathsf{T}}\right]^{-1}$$

where $\ell\left(\theta \mid \widetilde{\boldsymbol{u}}\right) = \log L(\theta \mid \widetilde{\boldsymbol{u}})$. By considering the distribution of $\widehat{\theta}$ 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 poosterior distribution for θ given $\widetilde{\boldsymbol{u}}$, or

$$p\left(\theta \mid \widetilde{\boldsymbol{u}}\right) = kL\left(\theta \mid \widetilde{\boldsymbol{u}}\right) p(\theta), \qquad (2)$$

where k is a normalizing constant and $p(\theta)$ conveys knowledge about θ before a value of $\boldsymbol{U}$ 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 tan individual examinee's θ from this extended observation, assuming that both the IRT model and the item prarmeters are known. To this end, we adapt notation from Little and Rubin (1987) and Rubin (1976) in defining the following terms:

- $\boldsymbol{u} = (u_1, \ldots, u_n)^\mathsf{T}$ is the (hypothetical) random vector of responses to all items in the full item set.
- $\boldsymbol{m} = (m_1, \ldots, m_n)^\mathsf{T}$ 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.

- $\boldsymbol{v} = (v_1, \ldots, v_n)^\mathsf{T}$ 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 $\boldsymbol{m}$, say $\widetilde{\boldsymbol{m}}$, effects a partition of $\boldsymbol{u}$, u, $\boldsymbol{v}$, and v according to which elements are observed and which are missing. That is, we may write $\boldsymbol{u} = (\boldsymbol{u}_{\text{mis}}, \boldsymbol{u}_{\text{obs}})$ to distinguish the missing and observed elements of $\boldsymbol{u}$, respectively. Similarly, $u = (u_{\text{mis}}, u_{\text{obs}})$, $\boldsymbol{v} = (\boldsymbol{v}_{\text{mis}}, \boldsymbol{v}_{\text{obs}})$, and $v = (v_{\text{mis}}, v_{\text{obs}})$. As with $\widetilde{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{m}}$, let $\widetilde{\boldsymbol{v}}$ denote a realized value of $\boldsymbol{v}$.

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 $\widetilde{u} = (0, 1)$, $\widetilde{m} = (1, 0)$, and $\widetilde{v} = (0, *)$.

Inferences must of course be based on the data that are actually observed, namely realizations of $v = (j_{\text{obs}}, m)$. Modeling the hypothetical complete data vector (u, m)–e-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 $p(m \mid u)p(u)$ or as $p(u \mid m)p(m)$, 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 Equation (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 $(u_{\text{mis}}, u_{\text{obs}})$.
- $g_\phi(m \mid u)$ is the probability that $\boldsymbol{m}$ takes value $\boldsymbol{m} = (m_1, \ldots, m_n)^\mathsf{T}$ given that $\boldsymbol{u}$ takes the value $\boldsymbol{u} = (u_1, \ldots, u_n)^\mathsf{T}$ 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 $g(m \mid u, \theta, \phi)$ to emphasize the dependence on θ.
- $h_\theta(u \mid m)$ is the probability that $\boldsymbol{u}$ takes the value u given that $\boldsymbol{m}$ takes the value m.
- $t_\phi(m)$ is the probability that $\boldsymbol{m}$ 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

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

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

$$L(\theta, \phi \mid \widetilde{v}) = \delta\left[(\theta, \phi), \Omega_{\theta\phi}\right] \cdot \int f_\theta(u_{\text{mis}}, \widetilde{u}_{\text{obs}})\, g_\phi(\widetilde{m} \mid u_{\text{mis}}, u_{\text{obs}})\, \mathrm{d}u_{\text{mis}},$$

where δ takes the value 1 if a value (θ, ϕ) is in the parameter space $\Omega_{\theta\phi}$ and 0 if it is not. 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 $\widetilde{m}$ and $\widetilde{u}_{\text{obs}}$. Using local independence, we can bring the probability for the observed responses outside the integral:

References

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581