An analysis of different psychometric treatments of item nonresponse in the SACMEQ Reading test

Hamish Coates

Centre for the Study of Higher Education
University of Melbourne

This analyses the effects of unplanned item nonresponse on the measurement of student ability in large scale educational achievement studies. It explores variations in student and national ability estimate distributions as a result of different psychometric treatments of nonresponse. A range of replicated analyses are presented using data collected from the crossnational SACMEQ study of Reading achievement. The relative merits of these treatments are summarised in conclusion, and implications for normative and methodological research and practice are considered. It is suggested that environmental and psychological antecedents of nonresponse need to be determined and that related variables be considered during the production of individual and national estimates of achievement.

Unplanned item nonresponse in large scale educational assessment

This paper presents an empirical analysis of the effects and various treatments of unplanned item nonresponse on the measurement of student ability in large scale educational achievement studies. Such nonresponse occurs when students omit responses to multiple choice items in tests they were administered and expected to answer. Nonresponse is a widespread phenomenon in educational assessment, manifesting even in situations where test scores are calculated on the number of correct items, and where students are explicitly encouraged to answer every item. The presence of nonresponse raises a range of psychometric questions, and the incidental or deliberated responses to these can have significant implications on assessment approaches and outcomes. Rather than targeting peripheral issues, the questions raised by nonresponse challenge our understanding of how students interact with items, and of how we should produce individual and national achievement estimates of student achievement.

There is an intimate link between national policy-relevant estimates of educational achievement and an individual student's response to a series of test items. In addition to providing a correct or incorrect response to a test item, students also have the choice of providing no response at all. Such response decisions are influenced by a wide range of individual and environmental factors, including testing contexts, individual motivation and aptitude, item contents, cultural factors, and school and pedagogical issues. Students' interactions with test items are also affected by their perceptions of scoring procedures, procedures which also play a role in converting individual results into national achievement estimates. An important part of the validity of national estimates derives from the capacity of such psychometric procedures to accurately and efficiently capture the complex dynamics of individual performance, and to transform these into national achievement estimates.

As shown below, the treatment of unplanned item nonresponse can have significant implications on the quality of national estimates. Despite its significance, however, nonresponse has received only very limited attention in psychometric research literature. Item nonresponse is an understudied phenomenon compared with related issues such as guessing, differential item functioning and cheating. Although the issue received occasional attention as a by-product of formula scoring or guessing investigations in the 1970s (Lord, 1974, 1983; Choppin, 1974), there is a shortage of

studies investigating the sources of nonresponse in testing conditions that do not penalise incorrect answers. Reflecting such limitations, Keeves and Masters (1999: 270) note that

> *The practical issues involved in the treatment of missing data both in calibration and scoring have not been sufficiently explored for full resolution of the problems to have been achieved. This issue would seem of considerable practical significance for further work to be done until consensus is reached on the most appropriate procedures to be employed.*

A number of studies, such as Grandy (1987), Little and Rubin (1987), Mislevy and Wu (1988, 1996), Lietz (1996), Koretz, Lewis, Stewes-Cox and Burstein (1993), Gonzalez, Adams, Wu and Ludlow (1997) and Ludlow and O'Leary (1999), stand out as exceptions against this scarcity, and provide a basis for the current research. Nonresponse is surely a product of situational and psychological as well as psychometric factors. While it is unlikely that the psychometric perspective is sufficient to answer the uncertainties generated by nonresponse, it is very likely a necessary part of any response to the issue of item nonresponse.

The current analysis considers test item nonresponse as a form of measurement disturbance. Smith (1994a: 5410-11) defines measurement disturbances as "those things which interfere with the measurement process". Measurement disturbances can be either attributed to persons, to items, or to interaction between persons and items (Smith, 1985). Adams and Wright (1994: 246-49) relate these disturbances to specific response manifestations such as "startup-anxiety, plodding, sloppiness, item-person interaction, item bias, item multidimensionality, copying, guessing, fatigue, calibration noise and random misfit". Measurement disturbances confound the assessment process to the extent that they mediate students' test performance or the subsequent estimation of their ability. Nonresponse can be considered to be a 'catastrophic measurement disturbance' which contests the limits of psychometric model functionality. Mislevy and Wu (1996: 1), for example, contend that "when incomplete data are encountered, the model that determines responses is embedded in a more encompassing model that determines which responses will be observed and which will be missing". Given its potential implications and unintended pervasiveness in educational testing, nonresponse is a notable absence from these and other measurement disturbance discussions.

The phenomenon of unplanned item nonresponse raises a number of questions: How can nonresponse itself be measured? What sort of implications does nonresponse have for national achievement estimates? Can nonresponse be analysed and treated psychometrically? How does nonresponse affect the psychometric estimation of student ability? Is nonresponse psychometrically ignorable, or does it need to be considered explicitly in ability estimation? What factors need to be considered in a more general psychometric model of response behaviour? While these questions are used in this paper indirectly to guide discussion, the present analysis is focussed on one question in particular: What are the effects of nonresponse and various psychometric treatments of nonresponse on estimates of student ability? Through an analysis of item nonresponse on the SACMEQ crossnational test of Reading ability, this study defines and tests a series of psychometric nonresponse treatments, and examines the consequences of their application for large scale educational measurement.

Identifying the nature and significance of unplanned item nonresponse

Distinguishing unplanned nonresponse from structural missing data

Contemporary large scale achievement studies often involve either planned or structural forms of missing data. Certain forms of planned or structural missing data can increase the quality and efficiency of measurement. Using item response modelling, for instance, student ability can be measured along a common construct using different but equated items. Targeted and adaptive

testing, the use of matrix item sampling and alternate test forms (Wright & Stone, 1979; Adams & Gonzalez, 1996; Mislevy & Wu, 1996) are examples of such applications in which intended missing data is introduced deliberately into a study and is controlled and interpreted by the analyst. Other forms of structural missing data may be detected post hoc, after testing has been conducted. This includes instances in which missing data is caused by accidental but still organisational or structural factors such as printing mistakes, translation problems, distribution of incorrect test booklets and page omissions. While this missing data is unintended, it is attributable to structural factors rather than students. It is possible that these factors can be used to account for the missingness during data collection, cleaning and analysis.

This paper focuses on nonresponse that originates from students. Student nonresponse is distinguished from structural missing data by an important logical difference. While structural missing data can be linked to the assessment design or organisational factors, student nonresponse is unplanned and occurs incidentally within the assessment process. Rather than being part of an assessment plan, it appears after the data has been collected. While it would be ultimately desirable to prevent student nonresponse by altering administration procedures or altering the design of instruments, the nature of the phenomena means that, except in highly controlled environments, this would be difficult if not impossible to achieve. The precise reasons for student nonresponse are often unknown, and psychological or contextual causes may even be paradoxical and resist easy explication. Unlike structural missing data, student nonresponse introduces unintended and, possibly, unidentifiable variation into the data, and hence confounds rather than clarifies assessment.

Identifying forms of unplanned item nonresponse

It is useful for analytical purposes to identify different forms of unplanned nonresponse. Researchers have identified two mutually exclusive varieties. Item nonresponse is distinguished as occurring because items have been skipped, or because they have been unreached. This distinction has been used in many contexts (Lord, 1980; Koretz et al, 1993; Mislevy & Wu, 1996; Gonzalez et al, 1997; Yamamoto & Kulick, 1999; Allen, Johnson, Mislevy & Thomas, 1999; Ludlow & O'Leary, 1999; Keeves & Masters, 1999; Allen, Carlson, Johnson & Mislevy, 2001).

Unreached nonresponse (R_U) appears at the end of a student's recorded responses and runs through to the last item on the test. This definition assumes that students have undertaken the test sequentially, have provided a response to at least one item and have not interacted with the unreached items. This definition implies that missing data for the last item is always classified as unreached nonresponse, that students who have not provided a response to at least one item are considered absent from the test, and that students who do not reach a particular item have not, by definition, attempted the remaining items. Unreached nonresponse can be summarised sufficiently by a single number. If k represents the quantity of unreached nonresponse in a test with L items, then there are $0<k\leq L-1$ amounts of unreached nonresponse. This definition differs slightly from the TIMSS 1995 (IEA, 1996) position that the first one or two unanswered items should be treated as skipped rather than unreached items (Gonzalez et al, 1997).

Skipped nonresponse (R_S) is linked to omissions that occur before the last answered item. This definition of skipped nonresponse is based on the assumptions that students have worked sequentially through a test from start to finish, interacted with the items that they have decided not to answer and, at a minimum, encountered at least two items and responded to only one. For a test with L items, and with $L\geq 2$, there are $L-1$ different amounts of skipped nonresponse and $\sum_{i=1}^{i=L-1} C_i^{L-1} = 2^{L-1}-1$ different combinations of skipped nonresponse. If a test varies in length between k=2 and k=L items due to unreached nonresponse, there are

$\sum_{k=2}^{k=L}\sum_{i=1}^{i=k-1} C_i^{k-1} = \sum_{k=2}^{k=L} 2^{k-1} - 1 = 2^L - 1 - L$ possible total nonresponse (R_T) patterns. This study focuses on the level or quantity rather than the location of skipped nonresponse. That is, no differentiation is made between skipped nonresponses located at the start, centre or end of a test, the type of items which were skipped, or between skipped nonresponses that occur in isolation or groups. The generality of the current definition incorporates unusual situations such as when only the last item is answered yet all preceding nonresponses are labelled as skipped. The same level of skipped nonresponse may carry different connotations depending on which items are skipped, where the skipped items are positioned in the test and how they are positioned in relation to each other.

Distinguishing the statistical implications of nonresponse

Nonresponse has different statistical implications depending upon its prevalence and distribution within a data set. Rubin (1976, 1987) and Little and Rubin (1983, 1987) initiated a general means of classifying the nature and implications of missing data. Specifically, missing values of a random variable can be missing completely at random, missing at random, or nonignorable. When data is missing completely at random, missing values approximate a small completely random sample from all values and are not stratified by any manifest or latent variable. As such, the missingness can be considered as independent to the distribution of observed responses. Nonresponse may therefore be ignored in estimation without affecting the accuracy or efficiency of estimates. When data is missing at random, in contrast, distributions of missing and observed data are only conditionally independent. Given suitable specification of a set of partitioning variables, missing values can be stratified and controlled according to particular covariates. Such variables could be identified by exploring contextual patterns in missing data using collateral information collected during achievement testing. When data is nonignorable, the missingness is considered to be nonrandomly related to and hence confounded with a variable under study. Alternatively, the distribution of missing data is related to an unobserved variable not included in the study. In this instance, there is insufficient information to distinguish missing data patterns from performance on the variable being measured. These classifications provide a general set of distinctions useful for clarifying and analysing the implications of and the suitable reactions to missing data.

Rubin's taxonomy has been applied previously to the interpretation of achievement test data. In studying the implications of nonresponse on maximum likelihood and Bayesian estimation procedures, Mislevy and Wu (1988, 1996) contend that skipped nonresponse is nonignorable. Mislevy and Wu (1996: 33) note that

> *When examinees are presented items, have 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, as sketched in Lord (1983).*

Mislevy and Wu (1988, 1996) suggest that if speed and ability are independent, unreached nonresponse may be statistically ignorable. Research into the relationship between response latency and ability, however is inconclusive (Grandy, 1987; Mislevy & Wu, 1988; Jansen, 1997). In general, therefore, the consequences of skipped and unreached student nonresponse on ability estimation cannot be ignored. That is, rather than analysing response data directly as received in the item-level data file, it is necessary to generate a psychometric approach to the interpretation of nonresponse. Failure to properly mitigate the influence of nonresponse may generate substantial biases and inefficiencies in ability estimation.

Psychometric treatments of nonresponse

General approaches for treating nonresponse

A range of statistical and psychometric styles of missing data treatment have been developed. One way of treating nonresponse is to assume a lack of relationship between missing and recorded observations. The primary means by which this approach is operationalised is by using listwise or pairwise deletion of missing data (Beaton, 1994). Listwise deletion involves removing from analysis students with missing observations to any variables. While this removes the apparent effects of missing responses from the data set, it also reduces the number and type of observations available for analysis. Depending on the distribution of missing data, listwise deletion can reduce estimate accuracy and efficiency. Specifically, when missing data is not missing at random, but rather has systematic patterns which are confounded with a variable under study, listwise deletion may result in biased group estimates with inflated variances caused by sample size reductions. Pairwise deletion is a more sophisticated algorithm in which each student's responses are included or excluded in estimation on a variable by variable basis (Rubin, 1987). Pairwise deletion introduces ambiguities into estimation, including unknown group biases to the extent that missing data patterns are nonignorable. In psychometric contexts, the deletion treatments of missing data also impose obvious limitations on the possibility of individual reporting.

A second more complex but sensitive style of missing data treatment is to employ a method that imputes values in place of omitted responses. Such treatments are divided into procedures which relate observed and missing observations through single or multiple imputations. Marginal and conditional mean substitution, conditional mean substitution with error, and hot deck imputation are four dominant forms of single imputation (Little & Rubin, 1987; Rubin, 1987; Beaton, 1994). Mean substitution involves replacing missing data for each variable with the sample or subsample mean. While it is possible to localize the cases from which the means are computed, the approach excludes much information returned in observed data and assumes a strong uniformity and similarity across each variable between the missing and observed data. While mean substitution avoids deleting subjects from analysis, it can bias variable distributions according to observed responses and influence their kurtosis and variance. The most direct extension of mean substitution involves specifying a linear regression equation to impute a mean value conditioned on selected psychological or contextual variables. While regression imputation increases the information extracted from other variables in the response matrix, it implies that unobserved values are a direct linear function of observed responses. Given the imputation of a single value for each nonresponse, a stronger assumption is made that the linear relationship is specified without error (Rubin, 1987). As such, no allowance for the stochasticity of the unobserved responses is made, and the variance of the imputed responses matrix is correspondingly underestimated. In consequence, the regression approach can be elaborated by selecting the conditional values from assumed homoscedastic and normal conditional distributions. Hot deck imputation is an alternative approach that "finds for each nonrespondent a matching respondent, where matching means close with respect to variables observed for both" (Rubin, 1987: 9). While hot deck imputation is considerably more informed by the observed responses than approaches such as mean substitution, it maintains the equation between observed and unobserved responses and hence underestimates response variability.

Multiple imputation procedures are the dominant means of treating test item nonresponse in contemporary large scale assessments. Such procedures factor uncertainty into the substitution of values for missing observations. They do this by "replacing each missing value by two or more imputed values in order to represent the [sampling, measurement and, importantly, nonresponse] uncertainty about which value to impute" (Rubin, 1987: vii). The technique results in the production of multiple data sets which can then be subjected to replicated analyses. The results of these replicated analyses can be aggregated to produce more robust and representative estimates than those delivered by single imputation techniques. Mislevy (1991) has developed the "plausible values" procedure as the most extensive psychometric elaboration of multiple imputation given data missing at random or missing completely at random. The plausible values procedure is designed to

provide accurate and efficient population rather than individual score distributions. Rather than imputing values for particular missing observations, plausible values procedures involve drawing multiple imputations from posterior distributions estimated by Bayesian procedures using collateral information collected alongside achievement tests (Mislevy, 1991). The plausible values technique is used extensively in contemporary large scale education assessments (Mislevy, Johnson & Muraki, 1992; Mislevy, Beaton, Kaplan & Sheehan, 1992; Adams, Wu & Macaskill, 1996; Thomas and Gan, 1997). Following Lord (1962), the plausible values treatment facilitates the production of population estimates without requiring complete student level response vectors. By this position, "obtaining a best estimate of the parameters of the underlying distribution can proceed directly from the imperfect response data, via an appropriate marginal procedure, without attempting to produce point estimates for individuals along the way" (Mislevy Beaton, Kaplan & Sheehan, 1992: 142). As such, however, it does not provide a solution to treating missing data at the individual level.

The treatments of item nonresponse investigated in this analysis

Six nonresponse treatments are analysed in this paper. The treatments are developed principally in reflection and extension of methods proposed by Grandy (1987), Ludlow (1994), Gonzalez et al (1997), Ludlow and O'Leary (1999). The present study examines these treatments in an item response modelling context and, in particular, with respect to the Rasch measurement model. The treatments analyse missing data, following Mislevy and Wu (1988, 1996), as if it were at best missing at random but most likely nonignorable. Table 1 summarizes the treatments, which are referred to as either incorrect (INC), untreated (UNT), position (POS), length (LEN), item (ITM) or guessed (GUE). All of the six treatments use particular reference information to apply single imputations of either a correct, incorrect or missing response. Unlike plausible values techniques which use various kinds of demographic or contextual information to guide imputation, imputation mechanisms in the current study are limited to individual or group level nonresponse variables. The LEN and ITM treatments use group level information to guide imputation, while the INC, UNT, POS and GUE treatments all use individual level nonresponse information.

Insert Table 1 about here

The INC treatment

Under the incorrect (INC) treatment, items students do not answer are scored as incorrect before estimating their ability parameter (θ_{INC}). This is an unconditional treatment of nonresponse that is not mediated by substantive, pragmatic or psychometric variables. This approach has been recommended and used by a number of researchers (Adams et al, 1996; Gonzalez et al, 1997; Yamamoto & Kulick, 1999; Ludlow & O'Leary, 1999; Afrassa & Keeves, 1999).

Treating nonresponse as incorrect makes a series of assumptions about students' response capability in relation to the variable that venture beyond what is given in the observed data. Scoring all nonresponse as incorrect makes the assumption that students have interacted with every item and have been unable to answer correctly those for which they have not responded. Despite ambiguities concerning the relationship between ability and nonresponse (van den Wollenberg, 1979; Grandy, 1987; Longford, 1995; Mislevy & Wu, 1996; Choppin, 1974; Lord, 1974, 1983; Longford, 1995), the INC treatment makes an assumption of inability rather than partial or even complete ability. Scoring unreached nonresponse as incorrect, for example, draws a psychometric link between factors such as fatigue and response latency, and the ability being measured by the test. Incorrect response, however, is frequently a product of partially correct rather than totally false knowledge (Lord, 1974, 1975; Masters, 1982; Budescu & Bar-Hillel, 1993). If this applies to absent in addition to wrong response, then the assumed equation of nonresponse as incorrect may be false. The assumption of inability is particularly strong in the case of unreached items which students may not have even encountered. Conversely, in applying a partial credit logic even to binary coded items, it could be argued, following Bock (1972), that nonresponse should be considered as a legitimate

dimensional response. From this perspective, while an incorrect response may reflect an engagement with the items, nonresponse implies a lack of capacity for cognitive interaction. From whichever perspective the above argument is advanced, however, it appears that scoring nonresponses as incorrect may introduce some muddle between test taking behaviours and measurement of the targetted ability.

The UNT treatment

Nonresponse can be ignored or left untreated (UNT) for the estimation of ability. This approach resembles the pairwise treatment of missing data in that students' responses can be entered into analysis as they occur. Lietz (1996: 399), for example, concludes from an investigation of different approaches to scoring nonresponse in Rasch modelling contexts, that "no evidence emerged to preclude… [the] ignoring of omitted or not-reached items for scoring". From this perspective, students' ability estimates (θ_{UNT}) are based solely on provided responses. Items without responses are treated, effectively, as if they were not administered.

While Mislevy and Wu (1988, 1996) indicate that ignoring nonresponse is analytically incorrect, it may be justified on alternative grounds when a conservative analytical approach is desired which minimizes assumptions made about the missingness process. The treatment implies that little information about students' ability is obtainable from their nonresponse. Consequently, if Lord's (1974) argument is extended from formula scoring contexts, it could be argued that little information is lost by removing nonresponse from analysis. Further, including information about item missing data may introduce confounding information into response data. If nonresponse is a function of offdimensional response behaviour, it may be beneficial to separate its effects from the observed data.

The POS treatment

The position (POS) treatment combines the INC and UNT approaches depending on whether nonresponse is skipped or unreached. For this treatment, skipped items are scored incorrect and unreached items are treated as omitted. Ability estimates based on the POS treatment (θ_{POS}) can then be computed. The POS treatment is consistent with the approach used in the National Assessment of Educational Progress (NAEP) studies (Mislevy, Johnson & Muraki, 1992; Allen et al, 1999; Allen et al, 2001).

Treating nonresponse according to its position within the response string moderates the alternatives presented above. Following Mislevy and Wu's (1996) argument that unreached nonresponse may be ignorable statistically, the approach moves towards factoring time or fatigue effects out of the assessment process by ignoring unreached nonresponse. Based on the assumption that students have not encountered unreached items, priority is placed on measuring students' responses to complete items rather than making assumptions about items with which they have not interacted. While it is assumed that students skip items due to inability, it is also assumed that unreached items provide little information about students' interaction with the trait being measured.

The LEN treatment

Group level information can be factored into the analysis of student nonresponse. With the length (LEN) approach, developed during analysis of (Southern Africa Consortium for Monitoring Educational Quality) SACMEQ test data (IIEP, 2000a), systemic rather than student level variations are used to define treatment criteria. Treating nonresponse as missing at random, information about the average number of items completed by students within a group can be used to calculate revised test lengths in terms of the number of items students would not be penalized for excluding. The number of nonresponses ignored in subsequent analysis would match these revised test lengths, and remaining nonresponse would be scored as incorrect. Student ability (θ_{LEN}) would then be estimated.

The validity and efficacy of the length nonresponse treatment rests upon successful negotiation of the adjustment criteria. This is particularly important given that the LEN treatment has the potential to penalize more highly performing candidates if items later in the test are ignored. In particular, therefore, the basis and amount of adjustment require careful and defensible qualification. Adjustments in this study are made at a country level, although it could be argued that they should be made at the regional or even school level. Thus, under the LEN treatment, all students within a particular country have the length of their tests revised. The extent of adjustment also needs qualification. The present analysis adjusts test lengths downwards from the end of the test by the number of items attempted within each group by 95% of the students. Setting the length reduction in light of the number of items attempted by 95% of the students represents a consensus between the levels of nonresponse in each country and the number of test items required to maintain the integrity and coverage of the test. The level has been set for exploratory purposes, and without the more rigorous forms of substantive justification which would be required in any empirical application.

The ITM treatment

The item (ITM) procedure modifies the LEN strategy by changing the scoring criteria. The LEN treatment emphasizes unreached nonresponse by scoring items backwards from the end of the test. As an alternative, items with the highest level of nonresponse within a particular group could be flagged to be ignored prior to estimating ability parameters (θ_{ITM}). The ITM alternative is different to the LEN procedure in that it reacts to both skipped and unreached nonresponse.

For the ITM procedure, the distribution of nonresponse in a particular country is first analysed to determine the number of items attempted by 95% of students. This number of items is set as the revised test length. All items set are ranked according to their level of total nonresponse. Items with the highest levels of total nonresponse are excluded until only the revised number of items remains. Excluded items are ignored, or treated as if they had not been administered. In analyses involving multiple groups, the particular items excluded may vary for each group.

The LEN and ITM procedures make group rather than individual level assumptions about nonresponse. These treatments use missing data patterns manifest at a group level as information about students' opportunity to learn via exposure in the curriculum or the nature of their response. When large numbers of students skip specific items, there is evidence that these items have been administered inappropriately for measuring a targeted ability. Similarly, when the level of unreached nonresponse increases at a group level, there is increased support for the assumption that the test designer may have administered more items than suitable for a particular group. Rather than adapting to the contingencies of individual student response patterns, these treatments accord more with a post hoc design revision undertaken, in light of the data, by the analyst. In consequence, there is a possibility that the better candidates may lose an advantage from rapid accurate work because others are given credit for some omitted items.

The GUE treatment

Omission and guessing have been discussed as inverse behaviours (Birnbaum, 1968; Choppin, 1974; Lord, 1975; Grandy, 1987; Mislevy & Wu, 1988, 1996). While such analyses have been undertaken conventionally in accord with formula scoring rubrics designed to penalize guessing behaviour, the analogy may be generalized to scoring models that only model correct responses. By treating nonresponse as inverse guessing, pseudo guessed values can be generated in place of nonresponse before ability estimation (θ_{GUE}). Missing data is thus substituted with either the correct or an incorrect response.

Following Mollenkopf (1960), Lord (1974, 1980, 1983) and Grandy (1987), the GUE strategy is based on the assumption that students would likely have guessed their response if they had provided a response to an omitted item. Lord (1980) argues for a slight variation on this approach, contending that only skipped rather than unreached nonresponse should be randomly imputed. The GUE procedure can be viewed as a modification to the imbalance of the INC approach in which students are penalized for nonresponse but not penalized for random response or guessing. Redressing this imbalance becomes increasingly pertinent if there are "some [education] systems that explicitly encourage guessing when an answer is unknown by a student" (Ludlow & O'Leary, 1999: 618). Key administration instructions given to students when collecting the data analysed in this study do not discourage guessing. Students were told to "Do your best to answer each question. Even if you are not sure about the answer to a question, please choose the answer that you think is the best one. Then move to the next question" (IIEP, 2000b: 2). This argument needs to be treated with caution, however, as it is based on the unsupportable premise that complete individual response vectors contain guessed or random answers. The GUE strategy may be more productively defended as an expression of the assumption that students could have at least provided random responses for the items they omitted. As noted by Lord (1980), however, this perspective implies the doubtful assumption that random imputation introduces valuable information into response. Optimally and conservatively, therefore, the procedure may perhaps be viewed as a less radical adjustment than the INC nonresponse treatment.

The most direct means of imputing guessed values into students' scores is to determine an adjustment constant by which to inflate the ability measure obtained after estimation. This procedure was used by Grandy (1987: 5), who states that "we estimated that one in five unanswered items would have been correct if the examinee had guessed correctly at random. Therefore, for every five items not answered, the examinee's score was raised one raw score point". In order to psychometrically model this nonresponse treatment, however, it is necessary to impute values for particular items and then use the imputed data to estimate student abilities, optimally using routines that control for offdimensional responses (Linacre, 2000). The present study modifies the GUE approach in this way.

It is necessary to specify how many items should be imputed with correct responses. It is impossible to define precisely the number of items a student would have guessed correctly. The objective probability that a student will correctly guess an item with m distractors lies somewhere within the range 1/m and 1/2. The lower bound likelihood of successful guessing 1/m is the most conservative estimate in terms of making the least assumptions about ability. Analysts have shown that guessing is more often an informed rather than completely random process (Lord, 1974; Cross & Frary, 1977; Bliss, 1980). It may be legitimate, therefore, to vary the probability of successful guessing upwards according to situational, student or test factors which may be thought to influence guessing behaviour. For current purposes, the conservative lower bound probability of 1/m is used consistently for all items and students. Students with k nonresponses were thus given k/m correct and (k(m−1))/m incorrect imputations. A value of m=4 was assumed despite variation in the number of distractors in some items.

A procedure for identifying which items are recoded with correct and incorrect responses needs to be specified. Alternative procedures can be classified into three general categories. First, imputation can be undertaken uniformly whereby for a given level of nonresponse k, the first or last k/m omitted items are imputed with correct alternatives. Second, a more sensitive alternative could factor collateral information into the selection of items for imputation. Third, a probabilistic sampling strategy could be specified for item selection. A random start systematic selection procedure (Kish, 1965) was used to operationalize the third approach in this study. Specifically, k/m correct imputations were made to every fourth item after a student's first skipped or unreached omission.

The conduct of the empirical analyses

Characteristics of the student sample and the SACMEQ Reading test

Data used for this study was obtained from the 2000 crossnational SACMEQ study of Reading achievement (IIEP, 2000a). The test was conducted to obtain policy relevant information for ministries of education in 15 southern African countries. The tests are primarily conducted as one component of a broader survey program, and are not intended to be high stakes for participating students. One country was purposively chosen from the international sample for the current analysis. The country was selected due to its large levels of nonresponse. SPSS (SPSS, 1999) was used to draw a random subsample of 3000 students from the country data

The data used in the current study was collected using a multiple choice power test of Reading ability. The test consisted of 83 multiple choice items, all of which were scored dichotomously. The majority of these items had four response alternatives, although items 7 to 12 had three and items 35 to 40 had two. Although the test was unspeeded, it was designed to be administered in around 60 minutes. Time limits were established on the basis of trial testing to both minimize the effect of time constraints on the typical student's response and maintain international standards (Garden & Orpwood, 1996; Adams & Gonzalez, 1996). The tests were designed for students around 10 or 11 years of age. Students undertook the test in school groups according to conditions set out in detailed administration manuals (IIEP, 2000b) which listed procedures and standards intended to promote cross-contextually invariant assessment.

The measurement model used to analyse student performance

The simple logistic Rasch model (Rasch, 1980) is used for item response modelling in this study. The model gives the probability of person n with ability θ_n giving response x_{ni} to item i with demand δ_i as

$$P(x_{ni} \mid \theta_n, \delta_i) = \frac{\exp(x_{ni}(\theta_n - \delta_i))}{1 + \exp(\theta_n - \delta_i)} \qquad [1]$$

where $x_{ni}=0$ for incorrect response and $x_{ni}=1$ for correct response. The expression $\exp(\theta_n-\delta_i)$ is the likelihood or odds of correct response. (If O_{ni} is the odds of correct response, B_n the ability of student n on the variable being measured, and D_i the difficulty of item i on the variable being measured, then $\theta_n-\delta_i$ is an additive expression of the log odds $\ln(O_{ni})$ of correct response given $\theta_n=\ln(B_n)$, $\delta_i=\ln(D_i)$, $\ln(O_{ni})=\ln(B_n)-\ln(D_i)=\ln(B_n/D_i)$ and therefore the multiplicative ratio $O_{ni}=B_n/D_i$.) The term "$\theta_n-\delta_i$" is commonly referred to as a "logit" in Rasch measurement contexts (Wright and Stone, 1979: 17). According to [1], if θ_n is less than δ_i, then the probability of a correct response is less than 0.5. If θ_n is equal to δ_i, the probability of a correct response is a 0.5. If θ_n is greater than δ_i, the probability of correct response is greater than 0.5. The articulation of the simple logistic model in [1] makes clear the assumption that the probability of response is an additive function of one person and one item parameter only. For current purposes, dichotomously scored items are parameterized in terms of their Thurstonian thresholds (Wright & Masters, 1982; Masters, 1988) which represent the location on the latent continuum at which there is a 0.5 probability of obtaining a correct response. Importantly, unlike two or three parameter psychometric models (Lord & Novick, 1968), the model does not parameterize item discrimination or the probability of chance response. The model states, following Wright (1987), that all items measuring a common construct discriminate equally and that guessing or random response is an offdimensional response contingency and an incidental rather than intrinsic item characteristic.

The approach used for generation of ability estimates

This study considers the effects of nonresponse on student rather than item parameter estimates. As such, it is important to control for or eliminate variation in item parameter estimates. Controlling for

such variation generates the possibility of establishing a qualitatively and quantitatively invariant measurement environment for the subsequent analysis of student ability.

Rasch modelling enables the estimation of item parameters that are invariant across samples drawn from the same population. This between sample measurement invariance is derivative of the Rasch model's operationalisation of the principles of additive conjoint fundamental measurement (Wright, 1967; Wright & Masters, 1982; Engelhard, 1992). Such measurement invariance, however, is conditional upon both the targetted trait maintaining normative invariance within the population (Taris, Bok & Meijer, 1998) and the selected items having measurement properties which accord with the Rasch model (Wright & Masters, 1982). Given the presence of these mediating factors, and hence the contingent nature of invariance, it is necessary to establish empirically the invariance of a particular item set. The empirical invariance of item parameters can be established either post hoc by using a range of techniques developed for the analysis of differential item functioning (Adams & Rowe, 1988; Scheuneman & Bleistein, 1999) or a priori by enforcing equality through anchoring item parameters during the replicated estimation of student ability. Following Ludlow (1994), Gonzalez et al (1997) and Yamamoto and Kulick, (1999), the latter alternative is used in current analyses. This anchoring process enforces the dimensionality of the construct being considered.

A reduced data file was used to calibrate items and produce an item anchor file. Initially, listwise deletion was used to remove students with missing data for any variable from the data. This produced a considerably reduced data file with only 718 students. While the listwise deletion process considerably reduced the representativeness of the sample used for item calibration, it removed the implications of high levels of unreached nonresponse on the estimation of item parameters (Ludlow & O'Leary, 1999). The capacity of the Rasch model to calibrate sample independent item estimates, discussed by Wright (1967), suggests that such reductions in sample representativeness would likely not have implications on item estimation. Unlike Adams, Wu and Macaskill (1996), no random subsamples were drawn to produce further reduced data sets representative of the complete original sample. Rather, the listwise reduced data file was used in the estimation of item parameters. The Quest program (Adams & Khoo, 1993) was used to run the routines. The variable map shown in Figure 1 relates the distributions of item and student estimates. The map shows that both distributions are normal in shape and reasonably well centred or targeted on each other. The variance weighted mean square fit or "infit" statistics (Wright & Masters, 1982; Smith, Schumacker & Bush, 1998) of items varied between 0.87 and 1.17, suggested a high level of accordance with the Rasch model. These calibrated item estimates were then used to anchor all subsequent estimations of student ability.

Insert Figure 1 about here

Analysis of bias in ability estimates

One way of investigating the qualities of ability estimates produced under each treatment is by comparing them against a baseline condition. The baseline condition needs to provide stable reference limits in the face of nonresponse against which the treatments can be evaluated. The Rasch model relates raw scores with student ability via a students' proportion of correct responses. A students' ability logit is a function of their raw score r and the number of items L used for ability estimation. If L is decomposed as L=r+q+k, where q is a student's incorrect responses and k is the level of nonresponse, then the proportional relationship linking raw scores and ability estimates is seen to be not solely dependent on the raw score r or number of incorrect responses q, but also on L–k, the number of items answered. Nonresponse introduces an indeterminacy into the relationship between raw scores and Rasch model estimates. The effect of k>0 can be thought of as a post hoc reduction in test length. This indeterminacy is shown in Figure 2 which plots raw scores against Rasch logits ($\theta-\delta$). Figure 2 shows eight ogives, corresponding to the targeted test lengths L=83 and a series of reduced test lengths from L=73 to L=13. The ogives show that reducing item

numbers inflates ability estimates independent of any raw score increase. A baseline condition for the evaluation of nonresponse needs to be invariant with respect to such changes.

Insert Figure 2 about here

The number of relationships between raw scores and ability estimates implied by each nonresponse treatment can be determined. The number of relationships in the UNT, POS, LEN and ITM treatments varies with skipped and unreached nonresponse at individual and group levels. The INC and GUE treatments, however, involve imputing values for each item, even if no response was initially given. These imputations effectively eliminate nonresponse, and indicate that there would be a uniform relationship between ability estimates and the scores used for estimation. In the case of the GUE treatment, however, this is because inflated scores are used. As considered later, the relationship is not uniform when GUE ability estimates are plotted against raw scores as initially obtained. Only with the INC treatment, therefore, does the ogive relating raw student scores with modelled ability estimates have a single monotonically increasing form. The INC treatment is thus set as a baseline in this study against which estimate distributions from the other treatments are compared.

The INC baseline condition can be used to produce a measure of estimate bias for each treatment. The estimate bias produced by each of the treatments can be defined as the difference between the ability estimate produced under the INC treatment and the estimate produced using each of the others. While the INC treatment is unbiased by definition, five biases can be computed, corresponding with the UNT, POS, LEN, ITM and GUE treatments. Adapting Taherbhai and Young (2001), these measures of treatment estimate bias (B_T) are given as

$$B_T = \theta_T - \theta_{INC} \quad [2]$$

where θ_T is the estimate produced under one of the UNT (θ_{UNT}), POS (θ_{POS}), LEN (θ_{LEN}), ITM (θ_{ITM}) or GUE (θ_{GUE}) treatments, and θ_{INC} is the estimate produced using the INC method. A range of exploratory analyses of these statistics are undertaken below.

Results of the psychometric and statistical analyses

Distributions of item nonresponse

There was 6.38% skipped, 3.51% unreached and 9.89% total nonresponse in the data. Figure 3 breaks this down and shows for each item i the number of skipped, unreached and total nonresponses. While there is a moderate level of skipping behaviour for all items, five items had high levels of skipped nonresponse. Item 19 has three times more missing data than expected given distributions across the items as a whole. Items 35 to 40 were skipped by far more students than most other items. This increase coincides with a change in multiple choice response scale. Unreached nonresponse begins around item 40 and there is an approximately linear increase towards the end of the test.

Insert Figure 3 about here

A total of 2282 students omitted items. Most of these students omitted only a few items: 10.27% omitted one item, 7.33% two items, and 4.87%, 5.73%, 6.37%, 6.50%, 4.83%, 3.50% and 2.07% omitted three, four, five, six, seven, eight and nine items. 9.28% of students omitted between 20 and 41 items. Only 3.33% of students omitted more than half of the items.

Bias in ability estimates by nonresponse treatment

Figure 4 shows how raw scores r relate with Rasch ability estimates under each treatment (θ_{INC}, θ_{UNT}, θ_{POS}, θ_{LEN}, θ_{ITM} and θ_{GUE}). The graphs show large numbers of points falling below the ogive

in all but the INC treatment. As well as emphasizing different patterns between the raw and modelled scores, the plots show differences in the distribution variances.

Insert Figure 4 about here

Treatment estimate bias statistics (B_T) were computed for each student. Distributions of these are shown in Figure 5, which also summarizes the means ($M(B_T)$), medians ($MED(B_T)$), standard deviations ($SD(B_T)$), skewness ($S(B_T)$) and maximums ($MAX(B_T)$). Each treatment has different estimate bias implications given various kinds and levels of nonresponse. They vary, therefore, in the number of students with bias greater than zero ($N(B_T>0)$).

Insert Figure 5 about here

Figure 6 plots individual student estimate biases for the treatments (B_T) by total nonresponse (R_T). The plots bring out very different patterns in the concentrations and levels of estimate bias given different total nonresponse levels.

Insert Figure 6 about here

Analysis of the relationship between individual bias and nonresponse can be summarized by examining mean bias for each level of nonresponse. Figures 7 to 9 plot the mean bias of each treatment ($M(B_T)$) for skipped (R_S), unreached (R_U) or total (R_T) nonresponse. There is a general increase in mean bias with nonresponse, although this varies depending on the type and level of nonresponse. While students who omit no items obtain the score expected under the INC treatment, students who omit around 70 items can have their score inflated considerably. The amount of bias differs across treatments. The UNT treatment leads to the greatest bias for nearly all nonresponse levels. While the other treatments are fairly homogeneous, the GUE treatment becomes increasingly distinguished as nonresponse increases. Of the POS, LEN and ITM treatments, the ITM treatment in general shows the lowest bias levels.

Insert Figures 7, 8 and 9 about here

Differences in bias standard deviation conditional on total nonresponse levels (R_T) are shown in Figure 10. There are three shifts in bias distribution variation. While the bias variances are initially stable, there is a general increase towards the middle of the test and a third phase of large variation corresponding with high nonresponse levels. It is important to note that the variance fluctuations at higher nonresponse levels may partially be influenced by the low numbers of students.

Insert Figure 10 about here

Individual estimate biases and variance distortions have implications on country score distributions. While tested students are not missing observations from a sampling perspective, Figure 11 shows that individual nonresponse can bias aggregated results. Figure 11 summarizes country estimate distributions across the treatments. The means (M), variances (V) for each distribution are shown, and effect size indices give standardized differences between the means.

Insert Figure 11 about here

A review of the results of different psychometric treatments of nonresponse

A summary of the implications of the six nonresponse treatments

The distributions of estimates under each treatment are summarized in the raw score and ability estimate plots in Figure 4. As expected, the INC baseline treatment is related to raw scores as a

single monotonically increasing function of student ability. Figure 11 shows that scoring all omitted responses as incorrect in the presence of high nonresponse levels can negatively skew and decrease the median and means of national score distributions. Such deflation may be problematic in educational jurisdictions in which students are encouraged to provide deliberative responses rather than necessarily answer every item (Gonzalez et al, 1997). Imputation of wrong for nonresponse also leads to a more variable distribution, and can lead to mean estimate differences between the INC and other treatments of as much as one third of a standard deviation unit.

Using the UNT treatment and ignoring nonresponse during ability estimation produces the most inflated ability estimate distribution. Figure 4 shows a large scatter of points below the ogive relating the ability estimates of students with full item responses to their scores. This scatter could be expected given that the UNT treatment produces a raw score and ability estimate relationship for each level of unreached nonresponse. Correspondingly, Figure 5 shows that this treatment is linked with the highest bias means, maximums and number of students with inflated ability estimates. The bias grows consistently with nonresponse levels, as shown in Figures 6, to 9, however Figure 10 indicates that it is not the most unpredictable of the treatments. Figure 11 shows how the inflation increases the median at the country level, yet also leads to a range constriction in the group ability estimate distributions.

Figure 4 shows that the POS treatment has the next most random series of raw score and ability estimate relationships after the UNT treatment. This may be expected given that the POS treatment controls skipped but ignores unreached nonresponse. The treatment thus establishes a series of relationships between raw scores and ability estimates for each unreached nonresponse level in the data. Given the dominance of skipped nonresponse in the current data, however, controlling for skipped nonresponse enables the treatment to partial most of the bias out of students' estimates. Figure 5 shows that this treatment leads to the lowest number of students with estimate bias, the lowest mean and median levels of bias distribution, but also a very skewed distribution. While Figure 6 confirms this result, it also shows that the POS treatment can promote large outlying bias statistics given the presence of high levels of unreached nonresponse. This is manifest in Figure 8 which, when juxtaposed against Figure 7, shows the reactivity of the POS treatment given unreached nonresponse. Figure 11 shows that the skew introduced into the distribution of bias values inflates the variances of group score distributions.

The raw score and ability estimate ogives in Figure 4 suggest that while the LEN and ITM group level treatments place a limit on the upper bounds of the logit inflation, an envelope of inflated estimates is still present. Within this range, however, results in Figure 5 indicate that the ITM treatment is associated with lower bias levels, variances and numbers of students with bias. The plots in Figure 6 reinforces this trend across the range of total nonresponse values. Under the LEN treatment, there is a general smooth increase in estimate bias up until the level of adjustment followed by a decrease as any additional omissions are scored as incorrect. By only ignoring particular items, conversely, the ITM procedure does not promote such uniform patterns of inflation. These characteristics are reproduced in Figures 7 to 9. Figure 11 confirms the higher group medians and means of the LEN treatment. The more deliberate choice of items in the ITM treatment leads to more conservative estimate bias levels.

The GUE treatment has the most systematic form of estimate bias of the six treatments considered. Figure 4 reflects the proportional relationship between nonresponse and GUE imputation whereby students receive an extra correct response for every fourth item omitted. While Figure 5 shows that this treatment reduces the variance in most of the bias distributions, Figure 6 indicates that it promotes a number of outlying bias values. Despite this skewness, the mean as well as the median values in Figure 5 indicate that this treatment leads to the second lowest level of average estimate bias. While Figures 7 and 8 show that the GUE approach is the second most biased given high

levels of nonresponse, Figure 9 shows that it is the most consistent and predictable across the range of total nonresponse. By biasing the estimates of students with high nonresponse levels and correspondingly low raw scores, Figure 11 indicates that this treatment reduces the variance and increases the statistical efficiency of the overall group score distribution.

Conclusions and implications of the analysis

This study has presented an empirical analysis of the nature and treatment of test item nonresponse which originates from students and manifests to analysts after the collection of data. After defining two distinct kinds of nonresponse, six psychometric treatments were presented and the characteristics and consequences of these were considered. Empirical results have shown that the psychometric treatment of student nonresponse has implications on the production of individual ability estimates and on group score distributions. As expected when using proportional item response models, student ability estimates tend to be inflated with increases in nonresponse level. Each of the six treatments provides some means of controlling this inflation. While the INC treatment is the only alternative that maintains a singular relationship between raw scores and ability estimates, it also deflates ability estimates the most. Against this, the UNT and LEN treatments lead to the most inflated estimates, while the POS, ITM and GUE methods represent middle ground alternatives. The treatments also influence the variability of group score distributions, which may have particular consequences for subsequent statistical analyses and generalizations.

There are many ways in which the current set of treatments could be adapted or extended. The LEN treatment, for example, could be modified to reduce the test length within the prescribed group for all students and not just those who have omitted items. Similarly, the ITM treatment could be modified to exclude selected items for all students rather than those who have omitted particular items. That is, the ITM procedure factors characteristics of the observed data into a revision of the items considered as relevant for measuring a specified ability in a particular education community. Another way of adapting the current treatments is to draw them together into a more generalized composite treatment strategy. This approach may involve applying different treatments depending on student identification or the distribution of nonresponse in student or group data. Further treatments, perhaps extending the partial credit approach suggested by Bock (1972) could also be considered. As part of these generalizations it could be necessary to develop a more generalized definition of unreached and, in particular, skipped nonresponse that fits better with the ITM, LEN and GUE types of treatment.

One way of generalizing the psychometric treatment of nonresponse may be to link it with certain psychometric developments which move towards increasing the psychological sensitivity of item response models. In many respects, nonresponse is difficult to interpret beyond situations involving response constraints such as a penalty for guessing (Lord, 1975; Mislevy & Wu, 1996). While it is difficult to interpret nonresponse from a score maximization perspective, more general explanatory models may provide better interpretive mechanisms by factoring in affective variables like risk orientation and motivation, subjective response probabilities, and perhaps even educational and situational context. By venturing beyond models that parameterize subjects “strictly on the basis of propensities towards correct response” (Mislevy & Verhelst, 1990: 211), many response processes and behaviours once considered “irrational” (Lord, 1974, 1983) might be incorporated into an increasingly powerful measurement model (National Research Council, 2001). In order to better account for the presences of nonresponse, such analyses could venture beyond interpreting how an aggregation of items measures a targeted ability and develop questions about the structure and functionality of the integrated series of cognitive processes that compose this ability.

Strategies for treating nonresponse could also be embedded within more general psychometric and statistical models. The psychometric study of nonresponse could be situated within a series of

psychometric, psychological and sociological contexts. While the proximal antecedents of nonresponse may be primarily psychological, analyses with causative models could also investigate the extent to which these are mediated by and interact with psychometric and situational factors. Examining these interactions would involve extending the definition of the covariates of nonresponse outlined above and using covariance structure analyses (Bollen, 1989), to model their interactions. This type of study may require both qualitative and quantitative forms of analysis undertaken within an experimental rather than quasiexperimental study design. Developing such a model and showing it to hold cross-contextually would be an important move towards understanding the aetiology of nonresponse and hence refining the methods for its interpretation and treatment.

The methodology of the current study could be adapted and extended. There is a need for further exploration of the theoretical consequences of each of the treatments. This would likely involve tightening links between the scoring rules for unanswered data, estimation routines and the Rasch model. Also, while this study has focussed on analysing the treatments with reference to a large scale crossnational set of data, there is a need for more localized analyses of the type exemplified by Wright and Stone (1979) and Adams and Wright (1994). Refined a priori modelling exercises could be complemented by simulation studies, similar to those undertaken by Smith (1988, 1994a, 1994b, 1996) and Smith et al (1998), which would facilitate a more directed analysis of each treatments' behaviour in the presence of specific kinds of missing data distributions. The investigation could also be extended beyond the Rasch model to consider models which parameterize items using two and three parameters.

The findings presented above have a range of practical implications. They reinforce, for instance, the need to ensure that administration instructions are clear and made cross-contextually consistent. The instructions should be lucid and make less cognitive demand than the easiest test item. The instructions need to inform students of the treatment of nonresponse as well as right and wrong response. Additionally, there is a need to develop and include standards for the treatment of nonresponse in the design of crossnational studies. Among other things, these standards would need to specify procedures targeted explicitly to minimize nonresponse, minimum numbers and acceptable patterns of item response, and set criteria for the psychometric treatment of nonresponse. The empirical analyses of nonresponse in this study has suggested ways in which these ideas may be pursued.

References

Adams, R.J. and Gonzalez, E.J. (1996). The TIMSS test design. In M.O. Martin and D.L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report (Vol 1: Design and Development)*. Boston: TIMSS International Study Center.

Adams, R.J. and Khoo, S.T. (1993). *QUEST: The Interactive Test Analysis System*. Melbourne: Australian Council for Educational Research.

Adams, R.J., and Rowe, K.J. (1988). Item bias. In J. P. Keeves (Ed.), *Educational Research Methodology and Measurement: An International Handbook*. Oxford: Pergamon.

Adams, R.J. and Wright, B.D. (1994). When does misfit make a difference? *Objective Measurement: Theory into Practice Vol 2*. NJ: Ablex.

Adams, R.J., Wu, M.L. and Macaskill, G. (1996). Scaling methodology and procedures for the mathematics and science scales. In M.O. Martin and D. L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report (Vol 2: Design and Development)*. Boston: TIMSS International Study Center.

Afrassa, T.M. and Keeves, J.P. (1999). Changes in students mathematics achievement in Australian lower secondary schools over time. *International Education Journal 1(1)*, 1–21.

Allen, N.L., Johnson, E.G., Mislevy, R.J. and Thomas, N. (1999). Scaling procedures. In N.L. Allen and J.E. Carlson and C.A. Zelenak (Eds.), *The NAEP 1996 Technical Report*. Washington: National Center for Educational Statistics.

Allen, N.L., Carlson, J.E., Johnson, E.G. and Mislevy, R.J. (2001). Scaling procedures. In N.L. Allen and J.R. Donoghue and T.L. Schoeps (Eds.), *The NAEP 1998 Technical Report*. Washington: National Center for Educational Statistics.

Beaton, A. E. (1994). Missing scores in survey research. In T. N. Postlethwaite (Ed.), *The International Encyclopedia of Education*. Oxford: Elsevier Science.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Boston: Addison-Wesley.

Bliss, L.B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple choice tests using elementary school students. *Journal of Educational Measurement, 17(2),* 147–153.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37(1),* 29–51.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley and Sons.

Budescu, D. and Bar-Hillel, M. (1993). To guess or not to guess: a decision theoretic view of formula scoring. *Journal of Educational Measurement, 30(4),* 277–291.

Choppin, B.H. (1974). The correction for guessing on objective tests. *Evaluation in Education, 9,* 71–80.

Cross, L.H. and Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests. *Journal of Educational Measurement, 14(4),* 313–321.

Garden, R.A., and Orpwood, G. (1996). Development of the TIMSS Achievement Tests. In M.O. Martin and D.L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report (Vol 1: Design and Development)*. Boston: TIMSS International Study Center.

Engelhard, G. (1992). Historical views of invariance: evidence from the measurement theories of Thorndike, Thurstone and Rasch. *Educational and Psychological Measurement, 52*(2), 275-291.

Gonzalez, E.J., Adams, R.J., Wu, M. and Ludlow, L. (1997). Applications of item response theory in the third international mathematics and science study. In M. Wilson and G. Engelhard and K. Draney (Eds.), *Objective Measurement: Theory into Practice (Vol 4)*. London: Ablex.

Grandy, J. (1987). *Characteristics of Examinees Who Leave Questions Unanswered on the GRE General Test Under Rights Only Scoring. ETS Research Report 87–38*. Princeton: Educational Testing Services.

International Association for the Evaluation of Educational Achievement (IEA) (1996). TIMSS 1995. Boston: TIMSS International Study Center.

International Institute for Educational Planning (IIEP) (2000a). *Southern African Consortium for Monitoring Educational Quality (SACMEQ)*. Paris: UNESCO International Institute for Educational Planning.

International Institute for Educational Planning (IIEP) (2000b). *Southern African Consortium for Monitoring the Quality of Education (SACMEQ) Manual for Data Collectors*. Paris: UNESCO International Institute for Educational Planning.

Jansen, M.G.H. (1997). The Rasch model for speed tests and some extensions with applications to incomplete designs. *Journal of Educational and Behavioral Statistics, 22(2),* 125–140.

Keeves, J.P. and Masters, G.N. (1999). Issues in educational measurement. *Advances in Measurement in Educational Research and Assessment*. New York: Pergamon.

Kish, L. (1965). *Survey Sampling*. New York: Wiley Classics Library.

Koretz, D., Lewis, E., Stewes-Cox, T. and Burstein, L. (1993). *Omitted and Not Reached Items in Mathematics in the 1990 National Achievement of Educational Progress*. California: National Center for Research on Evaluation, Standards and Student Testing. (ED 378 220)

Lietz, P. (1996). *Changes in Reading Comprehension Across Cultures Over Time*. New York: Waxmann Munster.

Linacre, J.M. (2000). *Winsteps Rasch Model Computer Program*. Chicago: MESA Press.
Little, R.J.A. and Rubin, D.B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37*, 218–220.
Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
Longford, N.T. (1995). *Models for Uncertainty in Educational Testing*. New York: Springer-Verlag.
Lord, F.M. (1962). Estimating norms by item sampling. *Educational and Psychological Measurement, 22(2)*, 259–267.
Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39(2)*, 247–272.
Lord, F.M. (1975). Formula scoring and number right scoring. *Journal of Educational Measurement, 12(1)*, 7–11.
Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. NJ: Erlbaum.
Lord, F.M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48(3)*, 477–482.
Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Boston: Addison-Wesley.
Ludlow, L.H. (1994). Omitted and *Not Reached Items, Summary of the Technical Advisory Committee Meeting (Vol Appendix F)*. Boston: TIMSS International Study Center.
Ludlow, L.H. and O'Leary, M. (1999). Scoring omitted and not reached items: practical data analysis implications. *Educational and Psychological Measurement, 59(4)*, 615–630.
Masters, G.N. (1982). A Rasch model for partial credit scorings. *Psychometrika, 47(2)*, 149–174.
Masters, G.N. (1988). Measurement models for ordered response categories. In R. Langeheim and J. Rost (Eds.), *Latent Trait and Latent Class Models*. New York: Plenum Press.
Mislevy, R.J. (1991). Randomization based inference about latent variable from complex samples. *Psychometrika, 56(2)*, 177–196.
Mislevy, R.J. and Verhelst, N. (1990). Modelling item responses when different subjects employ different solution strategies. *Psychometrika, 55(2)*, 195–215.
Mislevy, R.J. and Wu, P.K. (1988). *Inferring Examinee Ability When Some Item Responses Are Missing*. Princeton: Educational Testing Service.
Mislevy, R.J. and Wu, P.K. (1996). *Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits and Adaptive Testing*. Princeton: Educational Testing Service.
Mislevy, R.J., Beaton, A.E., Kaplan, B. and Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29(2)*, 133–161.
Mislevy, R.J., Johnson, E. G. and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17(2)*, 131–154.
Mollenkopf, W.G. (1960). Time limits and the behaviors of test takers. *Educational and Psychological Measurement, 20*, 223–230.
National Research Council (2001). *Knowing What Students Know: The Science and Design of Educational Assessment.* Washington: National Academy Press.
Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: MESA Press.
Rubin, D.B. (1976). Inference and missing data. *Biometricka, 63(3)*, 581–592.
Rubin, D.B. (1987). *Multiple Imputation for Nonresponse In Surveys*. New York: Wiley.
Scheuneman, J.D., and Bleistein, C.A. (1999). Item bias. In J. P. Keeves and G. N. Masters (Eds.), *Advances in Measurement in Educational Research and Assessment*. New York: Pergamon.
Smith, R.M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45*, 433–444.

Smith, R.M. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement, 48*, 657–667.
Smith, R.M. (1996). A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement, 56(3)*, 403–418.
Smith, R.M. (1994a). Validation of individual test response patterns. In T. Husen and N. T. Postlethwaite (Eds.), *The International Encyclopedia of Education. Oxford: Pergamon*.
Smith, R.M. (1994b). Detecting item bias in the Rasch rating scale model. *Educational and Psychological Measurement, 54(4)*, 886–896.
Smith, R.M., Schumacker, R.E. and Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcomes Measurement, 2(1)*, 66–78.
SPSS. (1999). *SPSS Release 10.0.1*. Chicago: SPSS Incorporated.
Taherbhai, H.M. and Young, M.J. (2001). The impacts of rater effects on weighted composite scores under nested and spiralled scoring designs, using the multifaceted Rasch model. *Journal of Outcome Measurement, 5(1)*, 819-838.
Taris, T.W., Bok, I.A., and Meijer, Z.Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: a general approach. *Journal of Psychology*, *132*(3), 301–316.
Thomas, N. and Gan, N. (1997). Generating multiple imputations for matrix sampling data analysed with item response models. *Journal of Educational and Behavioral Statistics*, *22(4)*, 425–445.
van den Wollenberg, A.L. (1979). *The Rasch Model and Time Limit Tests*. University of Nijmegen: Unpublished Doctoral Thesis.
Wright, B.D. (1967). Sample-free test calibration and person measurement. Chicago: MESA Press.
Wright, B.D. (1987). Some comments about guessing. *Rasch Measurement Transactions, 1(2), 9*.
Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.
Wright, B.D. and Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.
Yamamoto, K. and Kulick, E. (1999). Scaling methodology and procedures for the TIMSS Mathematics and Science scales. *1999 TIMSS Technical Report*. Boston: TIMSS International Study Center.

Table 1. Summary of nonresponse treatments

Treatment	Summary
INC	Nonresponse is scored as incorrect.
UNT	Nonresponse is ignored and scored neither as correct or incorrect. Items with missing responses are treated as if they had not been administered.
POS	Skipped items are scored as incorrect and unreached items are ignored.
LEN	Ignores a determined number of item nonresponses backwards from the end of the test. Other missing responses are treated as incorrect.
ITM	Ignores nonresponses to a specific number of the least answered items. Other missing responses are treated as incorrect.
GUE	Responses simulating random guesses are imputed for missing data.

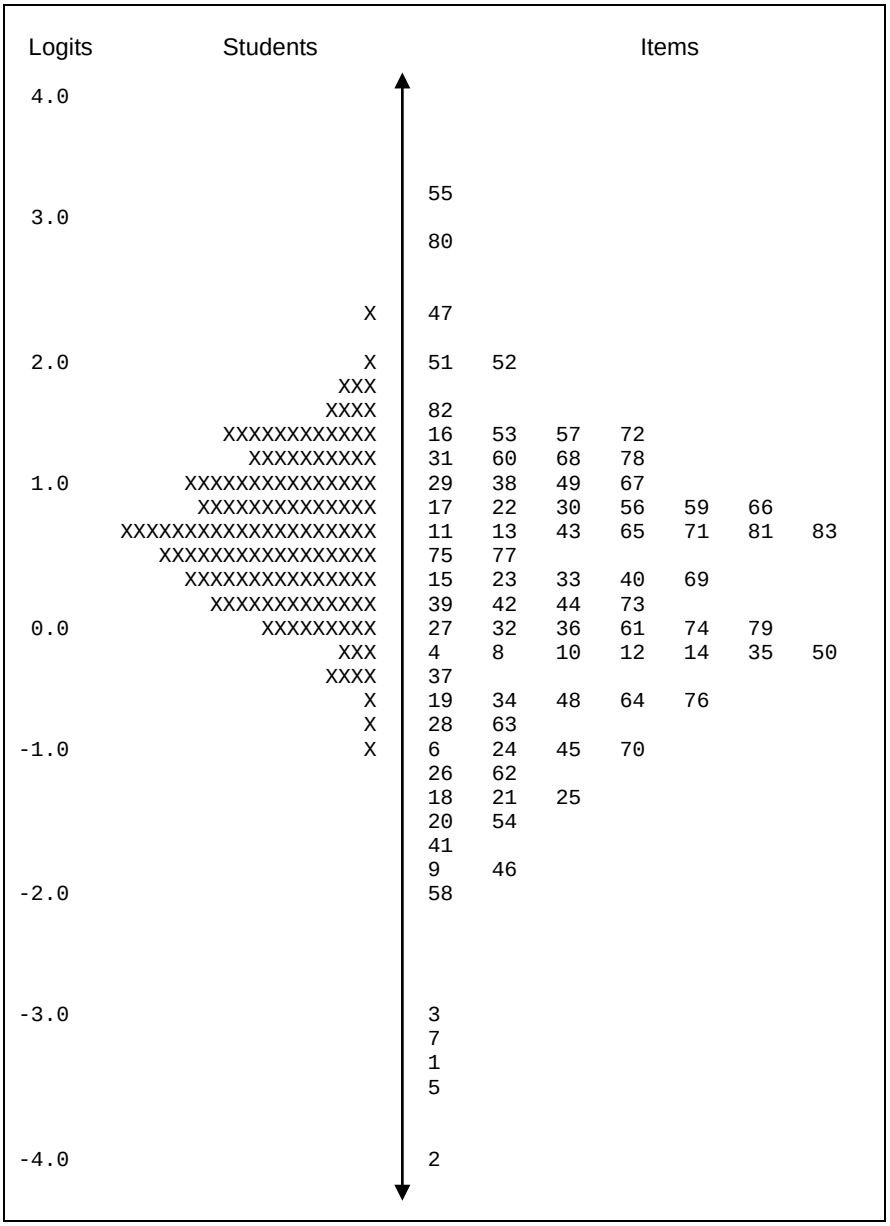


Figure 1. Variable map showing student ability and item estimate distributions.

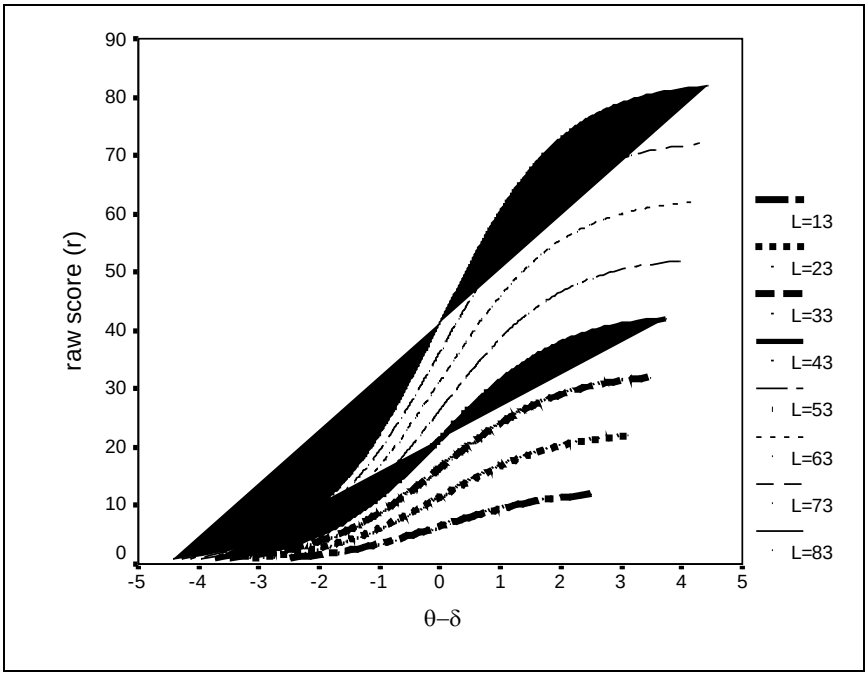


Figure 2. Raw score and ability estimate ogives by test length.

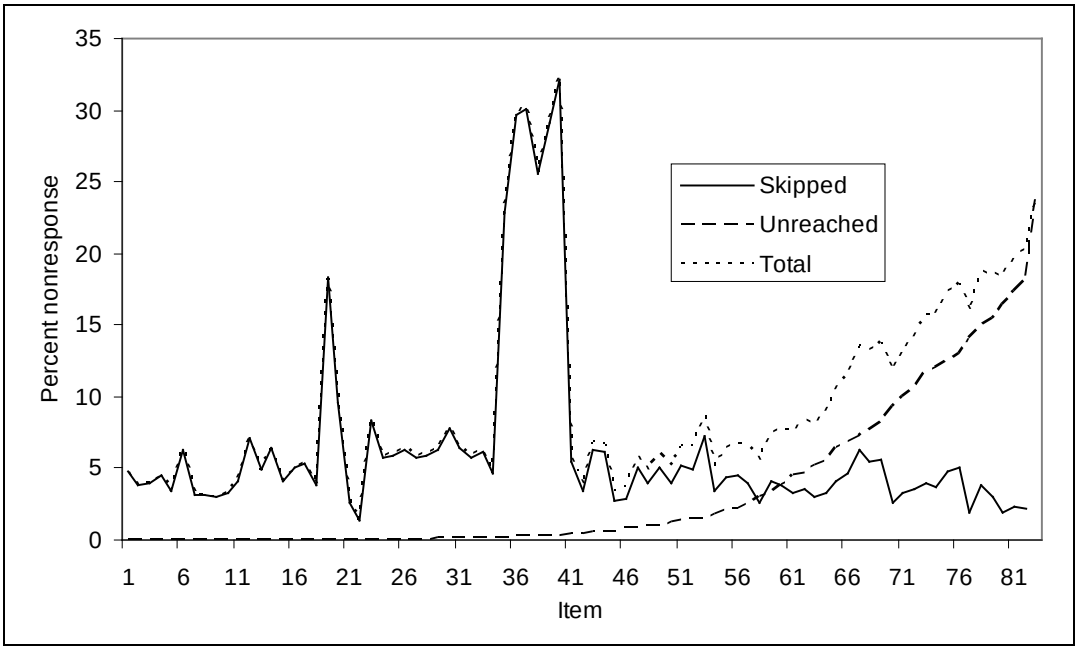


Figure 3. Percentage of skipped, unreached and total nonresponse by item.

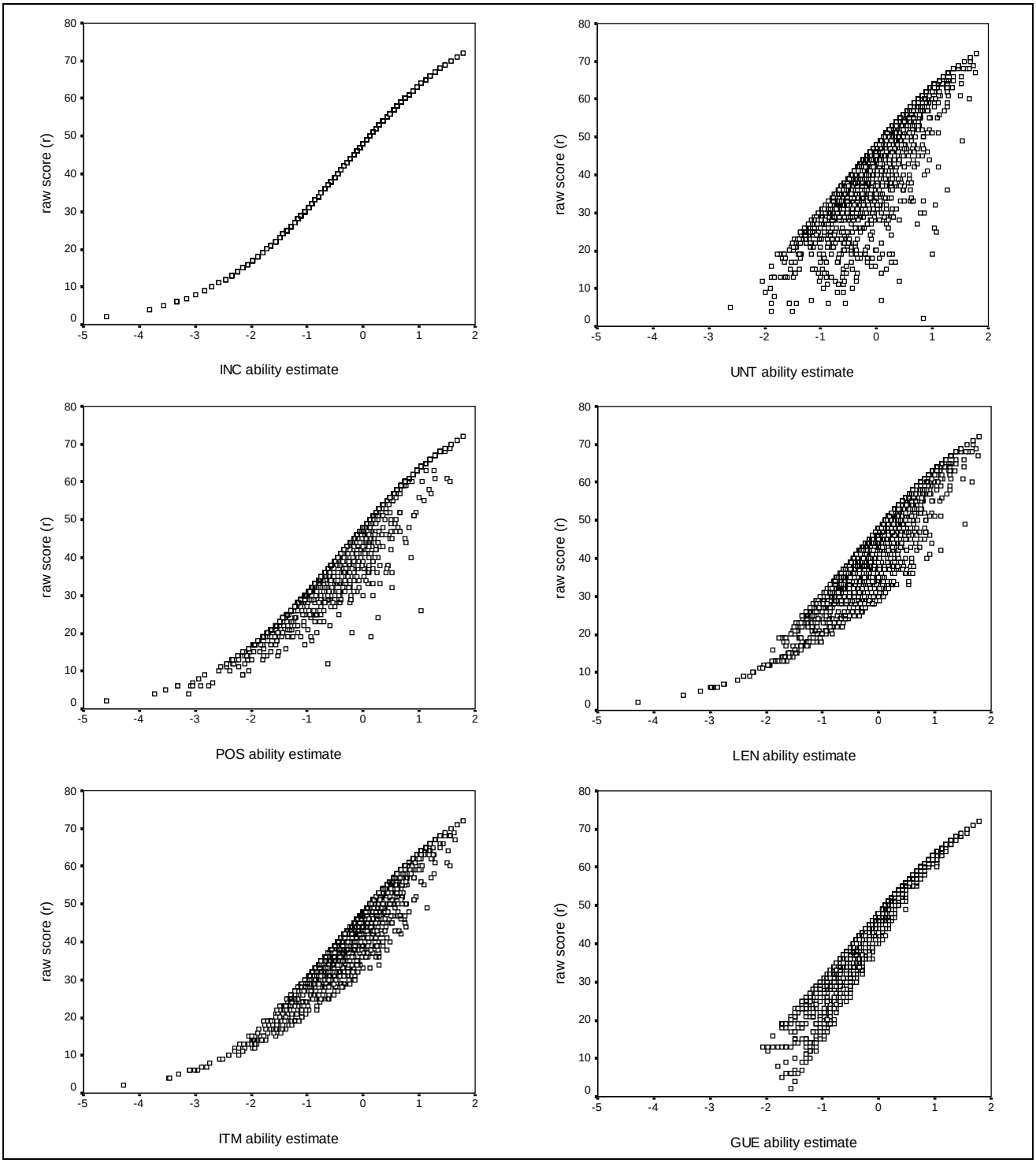
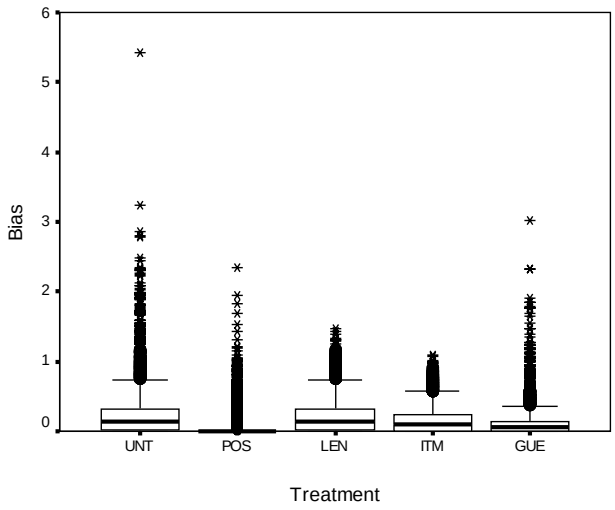


Figure 4. Raw score and ability estimate ogives by treatment.



	M	MED	SD
B_{UNT}	0.26	0.14	0.39
B_{POS}	0.07	0.00	0.19
B_{LEN}	0.22	0.14	0.26
B_{ITM}	0.16	0.09	0.19
B_{GUE}	0.14	0.06	0.23

	S	MAX	$N(B_T>0)$
B_{UNT}	3.33	5.42	2282
B_{POS}	4.14	2.34	718
B_{LEN}	1.55	1.47	2275
B_{ITM}	1.66	1.09	2104
B_{GUE}	4.08	3.02	1974

Figure 5. Distributions of estimate bias by treatment.

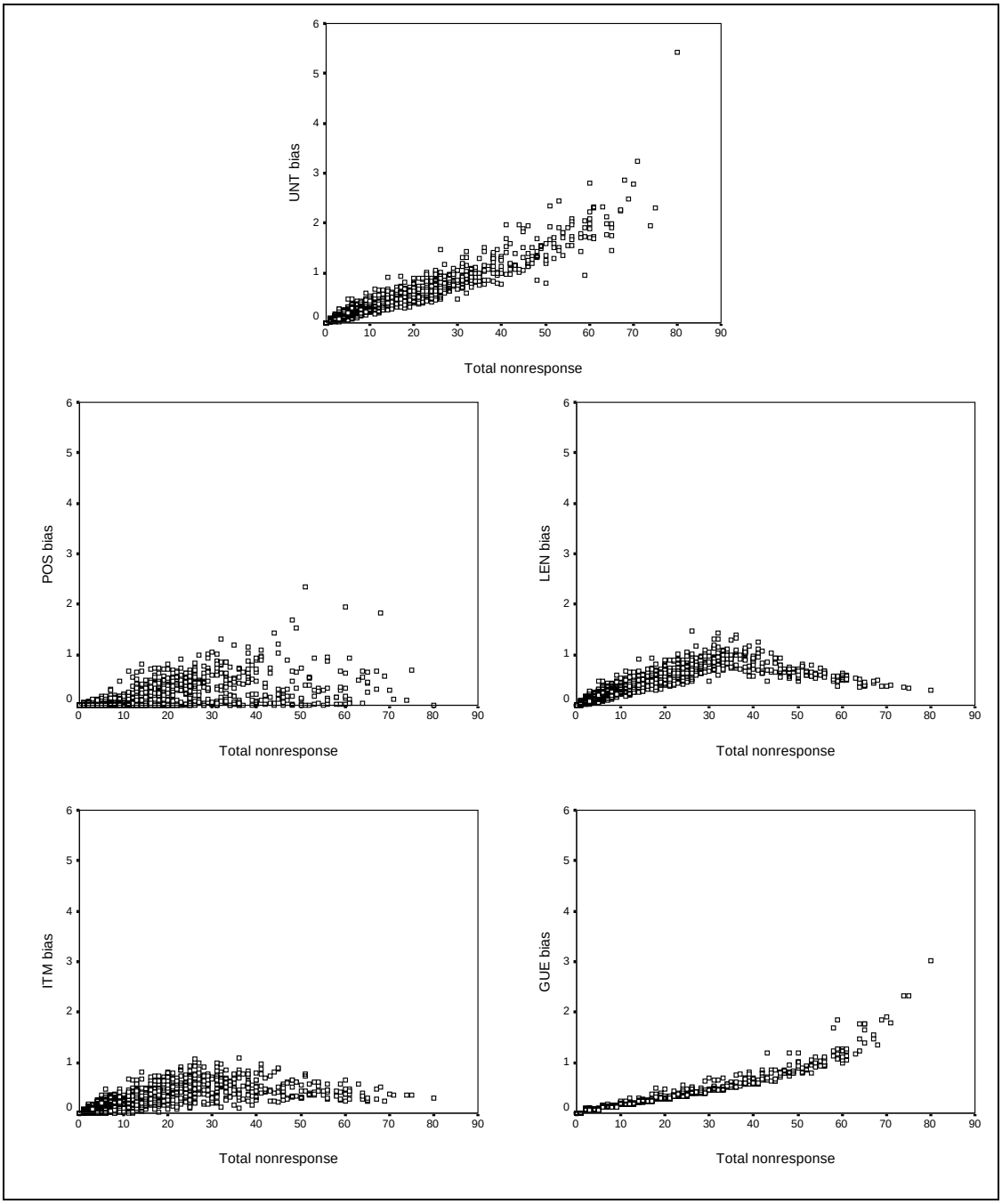


Figure 6. Estimate bias and total nonresponse relationships by treatment.

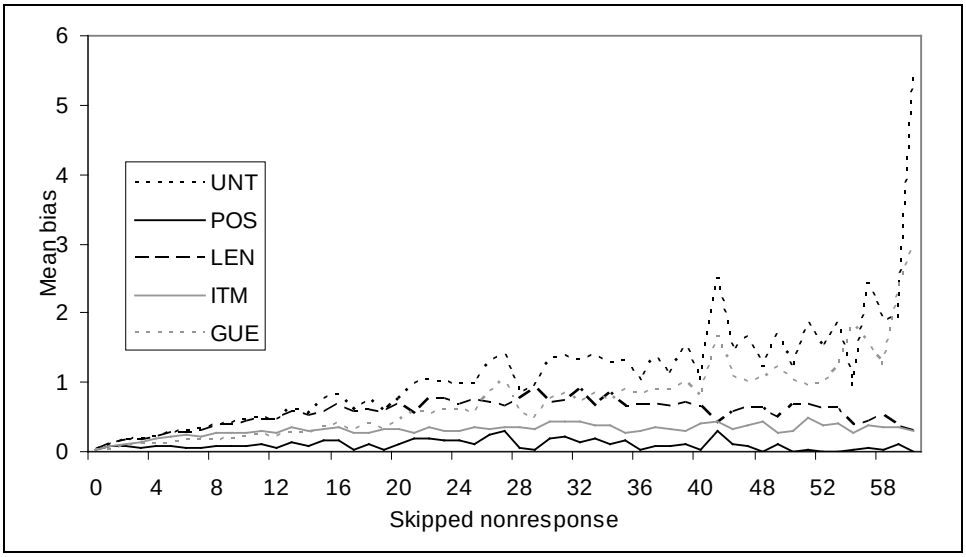


Figure 7. Mean estimate bias and skipped nonresponse by treatment.

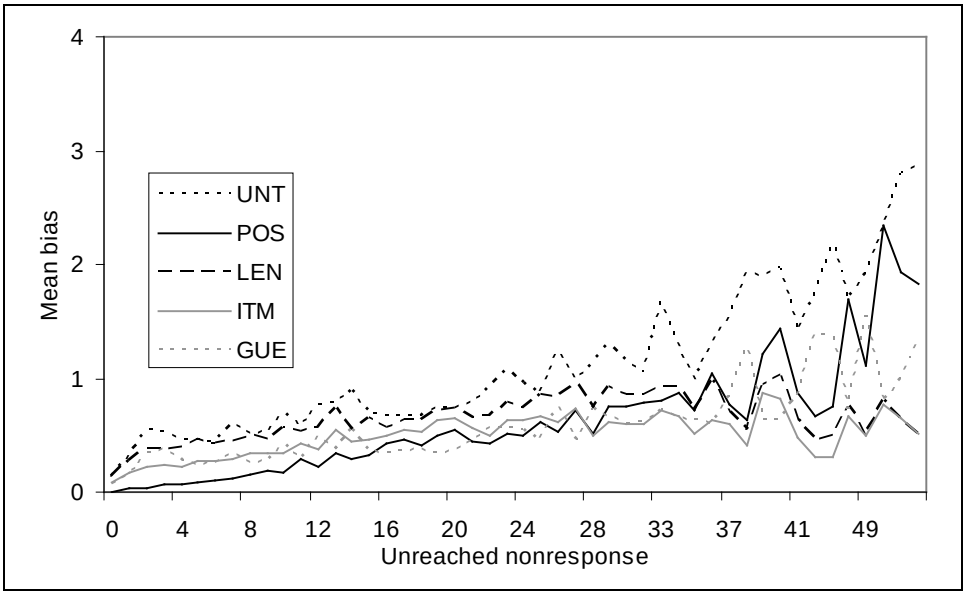


Figure 8. Mean estimate bias and unreached nonresponse by treatment.

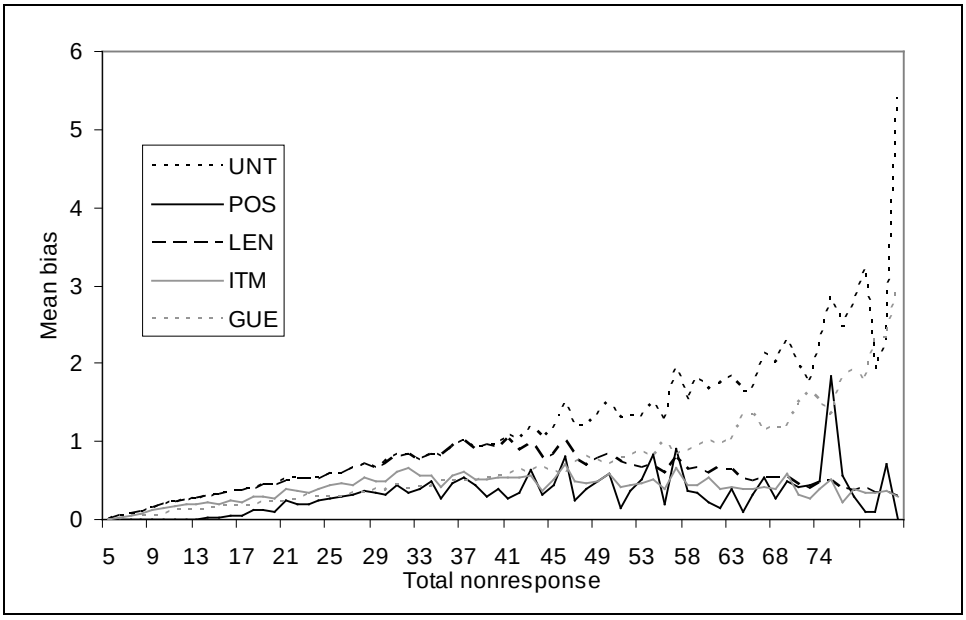


Figure 9. Mean estimate bias and total nonresponse by treatment.

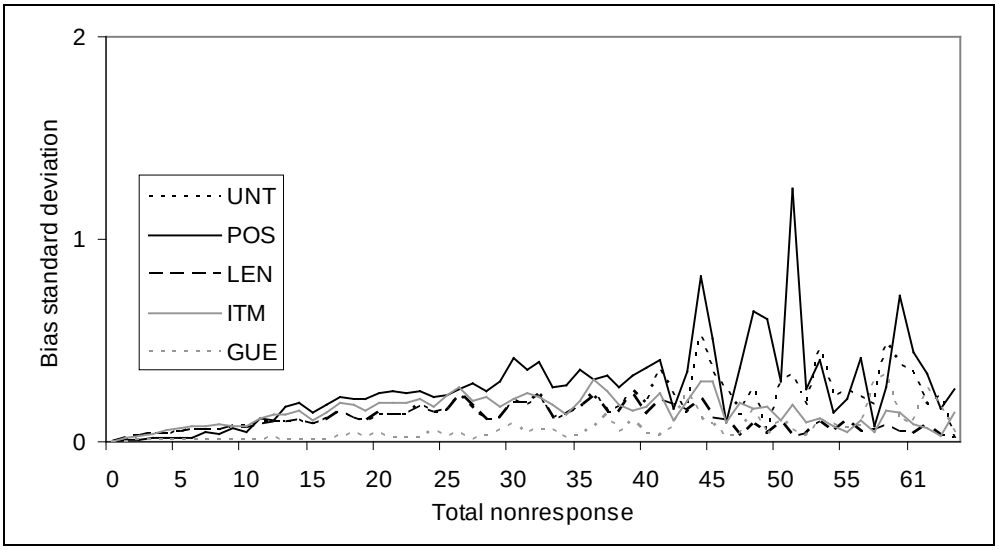
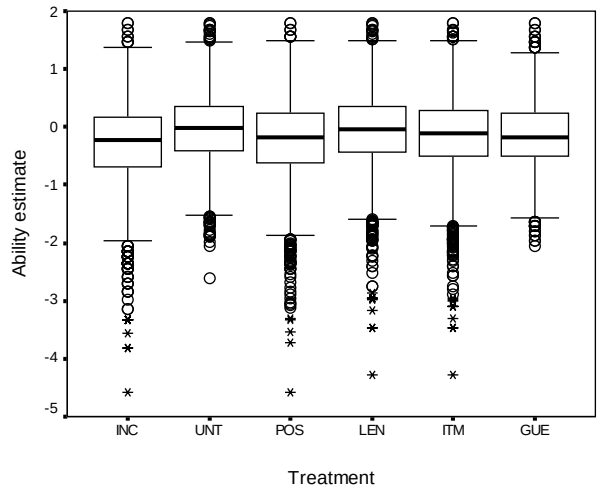


Figure 10. Estimate bias standard deviations and total nonresponse by treatment.



	M	V
θ_{INC}	–0.33	0.54
θ_{UNT}	–0.05	0.35
θ_{POS}	–0.25	0.49
θ_{LEN}	–0.10	0.42
θ_{ITM}	–0.16	0.46
θ_{GUE}	–0.18	0.36

	θ_{INC}	θ_{UNT}	θ_{POS}	θ_{LEN}	θ_{ITM}
θ_{UNT}	–0.31				
θ_{POS}	–0.08	–0.23			
θ_{LEN}	0.26	–0.05	0.18		
θ_{ITM}	0.18	–0.13	0.10	–0.08	
θ_{GUE}	0.16	–0.15	0.08	–0.10	–0.02

Figure 11. Country ability estimate distributions by treatment.