Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments

Eugenio Gonzalez
Educational Testing Service, Princeton, NJ, USA[1]

Leslie Rutkowski
Indiana University, Bloomington, IN, USA[2]

Large-scale assessments usually set out to cover an extensive content domain. Because of this, and to avoid overburdening students and schools, assessments are designed in such a way that each student is administered only a fraction of all the available items in the assessment. These designs are referred to as matrix sampling or multiple matrix sampling. This approach to test design and administration allows one to estimate sufficiently precise proficiency distributions of the target population and sub-populations and a complete coverage of the assessment framework, while reducing individual examinee burden and testing time at the school. Many of the choices and trade-offs in designing a multiple matrix sampled assessment are discussed and several example designs are described. A simulation study illustrates the impact that sparseness strategies can have on person and item parameter recovery. Implications for test design are discussed.

1 The opinions expressed herein are those of the author and do not necessarily represent those of Educational Testing Service.

2 The authors thank Andreas Oranje, Matthias von Davier, Rebecca Zwick, and an anonymous reviewer for suggestions and for comments on previous versions of this paper.

INTRODUCTION

Large-scale assessment (LSA) programs are charged with measuring what members of a given population know and can do in a given content domain or whether those members have acquired the skills necessary for performing future life activities. The breadth of topics measured by these programs is such that a large number of contents and skills are assessed. A number of national and international large-scale assessment programs exist. Each has its own focus and underlying philosophy. For example, the Trends in International Mathematics and Science Study (TIMSS) assesses mathematics and science knowledge and skills acquired by fourth and eighth graders (Mullis et al., 2005; Neidorf & Garden, 2004); the Programme for International Student Assessment (PISA) seeks to measure mathematics, scientific, and reading literacy of examinees who are 15 years of age (OECD, 2006); and the Progress in International Reading Literacy Study (PIRLS) aims to measure the reading literacy of fourth graders, an age when children are expected to make the transition from learning to read to reading to learn (Mullis, Kennedy, Martin, & Sainsbury, 2006). National large-scale assessment programs also aim to measure knowledge and skills of examinees of different ages and grade levels. For example, in the United States, the National Assessment of Educational Progress (NAEP) (National Center for Educational Statistics, 2010) assesses students at Grades 4, 8, and 12 in a number of content areas.

Because of the ambitious scope of these large-scale assessment programs, each assessment is designed in such a way that each student is administered only a fraction of all the available items in that assessment. In other words, each student is administered a particular combination of test items, thus ensuring sufficient content coverage across the population while reducing the assessment burden for any one student. The term multiple matrix sampling (Shoemaker, 1973), or, in older literature, item-sampling (Lord, 1962), arises from the practice of sampling both examinees and items; that is, giving samples of items to samples of examinees. Table 1 provides an example of multiple matrix sampling. In this example, each subject is administered a set of four of the six available items, and each item is administered to eight of the 12 examinees. Students 1 and 7 are each administered Items 1 through 4, Students 2 and 8 are each administered Items 2 through 5, and so on. Students 1 and 7 and Students 2 and 8, for example, may also be administered Items 2 through 4.

Matrix sampling of items is thus used in large-scale assessments to accommodate a broad coverage of the content domain, thereby ensuring that items are administered to a sufficient number of students without necessitating excessive testing time for any one individual student. Matrix sampling of items also allows us to estimate proficiency distributions of the population, while reducing individual examinee burden and testing time at the school, and representing the assessment framework satisfactorily. As is the case in most large-scale assessments, individual measurement precision is sacrificed in the interest of increased content coverage. This emphasis, or set of priorities, makes the booklet designs used in large-scale assessments rarely optimal for individual reporting, yet useful for group-level reporting. For operational expediency, large-scale assessment items are assigned to blocks that are then combined into forms

Table 1: Multiple matrix sampling

Subject	Set	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
1	1	x	x	x	x		
2	2		x	x	x	x	
3	3			x	x	x	x
4	4	x			x	x	x
5	5	x	x			x	x
6	6	x	x	x			x
7	1	x	x	x	x		
8	2		x	x	x	x	
9	3			x	x	x	x
10	4	x			x	x	x
11	5	x	x			x	x
12	6	x	x	x			x

according to a particular design or specification. This approach differs from other matrix sampling approaches where items are sampled randomly from the universe of possible items (see, for example, Barcikowski, 1972; Scheetz & Forsyth, 1977; Gressard & Loyd, 1991).

While mathematical optimization routines exist for optimizing booklet designs subject to a number of constraints (e.g., van der Linden & Carlson, 1999; van der Linden, Veldkamp, & Carlson, 2004), our focus in this article is not only on the practical issues that researchers need to consider when selecting a particular multiple matrix sampling design but also on the consequences of that process for the estimation of the population and item parameters. This paper contains two related sections. In the first, we briefly review the history, theory, and implementation of booklet designs. A substantial portion of this discussion includes details on current uses of matrix sampling in large-scale assessments (e.g., TIMSS, PIRLS, and PISA), the advantages and disadvantages of using multiple-matrix-sample booklet designs, and practical guidelines on selecting a particular booklet design. In the second section, we conduct an empirical investigation of the degree to which population and item parameters are recovered as a function of matrix sparseness and sample size during use of an incomplete booklet design. To explore this issue, we simulate data under different booklet designs. We then generate simulated responses to these booklet designs, assuming that a two-parameter logistic item response theory (IRT) model is sufficiently robust to assess the recovery of population and item parameters under different booklet designs.

THE ORIGINS OF MULTIPLE-MATRIX-SAMPLE BOOKLET DESIGN

Traditional approaches to sampling generally relied on sampling examinees from the population; however, a number of initial investigations (Johnson & Lord, 1958; Lord, 1962; Pugh, 1971) suggested that item sampling, in which small subsets of the total available items are administered in groups to the entire population, or multiple matrix sampling, could be an efficient and cost-effective way of assessing examinees and populations. Early work in the field of multiple matrix sampling, also referred to as item-examinee sampling and item sampling (Shoemaker, 1973), showed that, in many circumstances, this method is a reasonable (and sometimes advantageous) way to estimate group means (Gressard & Loyd, 1991; Johnson & Lord, 1958; Plumlee, 1964) and standard deviations (Gressard & Loyd, 1991; Pugh, 1971). Lord (1962) used empirical data to show that item sampling with replacement, where the entire normed population responds to small groups of possibly overlapping test items, is an effective means of recovering the known population norm values. Results from a study that mailed full and item-sampled questionnaires to a random sample of principals showed significantly higher response rates for the item-sampled questionnaires (Munger & Loyd, 1988).

A number of early empirical studies sampled items at random for assignment to subtests, or *forms* in large-scale assessment terminology (e.g., Lord, 1962; Plumlee, 1964; Shoemaker, 1970a, 1970b). During the same period, researchers used stratification techniques to assign items to subtests based on item difficulty (Kleinke, 1972; Scheetz & Forsyth, 1977) or item discrimination (Myerberg, 1975; Scheetz & Forsyth, 1977). Lord (1965) suggested that balanced incomplete block (BIB) designs might be advantageous for multiple-matrix sampling, because they fulfill the conditions that every item block appears an equal number of times in all block positions. This balancing of positions by means of BIB and variants of this design is a commonly used tool in experimental design (Nair, 1943; Yates, 1939). Knapp (1968) subsequently incorporated Lord's suggestion and found that a BIB design was an extremely efficient means of estimating the mean, variance, and reliability coefficient of several assessments. The BIB design was eventually, and successfully, implemented in the 1983/1984 NAEP assessment (Beaton, 1987; Beaton & Zwick, 1992; Johnson, 1992).

Many of the early studies depended on percent-correct methods for estimating test scores on multiple matrix sampled assessments (c.f., Johnson & Lord, 1958; Lord, 1962; Plumlee, 1964; Pugh, 1971). The rise of item response theory (IRT) methods in educational assessment facilitated the integration of multiple matrix sampling schemes (Bock, Mislevy, & Woodson, 1982). In this context, Mislevy (1983, 1984), Mislevy and Sheehan (1987), Reiser (1983), and others developed group-level models that estimated the underlying latent traits measured by assessments that followed a multiple matrix booklet design.

The dominant approach, applied to NAEP since 1983/1984 (Beaton, 1987) and clearly laid out in Mislevy (1991), uses a population model to integrate the advantages of multiple matrix sampling with IRT models. The approach is a latent regression IRT model. It integrates an IRT-based measurement model with a population model that utilizes covariates of proficiency. The latent regression model is utilized to estimate posterior distributions of examinee proficiency, given item responses and covariates. This posterior distribution, which draws on Rubin's (1976, 1978, 1987) multiple imputation technique, is used as the basis upon which to impute a set of plausible values for each examinee and each of the subscales.

Definition of Key Terms

Before proceeding further, we define some concepts in order to facilitate subsequent discussions. To understand the implementation of booklet designs, it is useful to clarify what is operationally understood by items, units, blocks or clusters, forms, and (ultimately) the assessment.

- *Items:* Items are the most basic unit of an assessment. Usually, an item is an individual task administered to a respondent, and it receives a score. Scores can be assigned automatically by machine or computer, or by people who examine the response and determine the best score based on the scoring rubric for the item. In general, when there is a single test form, scored item responses are summed for each examinee to compute a total score on a given test. Alternatively, methods that make assumptions about a latent trait (e.g., modern test theory methods) can be used to generate individual or group test scores. In a modern test theory framework, the score for each examinee is based on estimated item parameters. Even though observed correct scores can be used (after adjusting for form differences) to report results across multiple forms, modern test theory methods such as item response theory (IRT) are particularly useful when the assessment is composed of multiple forms. They are useful because they allow a formal evaluation of how well the different test forms can be aligned along a common scale.

  Items are generally of three kinds—multiple-choice, open-ended or constructed response, and performance. Multiple-choice items typically ask an examinee to read a given problem statement (usually referred to as the *stem*) and to select an answer from a fixed set of possible answers (usually referred to as the response *options*). The same item followed by the same response options, but in a different sequence, should generally be treated as a different item, unless there is evidence indicating that the order of the options does not affect the overall performance on the item. Open-ended or constructed-response items usually present a problem, and the examinee is expected to provide an answer in a specified format (e.g., essay or short-answer). Performance items pose a problem or question, and the examinee is asked to perform a task or demonstrate a skill (e.g., nursing students asked to measure a patient's blood pressure).
- *Units:* These consist of a common stem or stimulus that is followed by several items. An individual item is a special case of a unit. Examples of units include a reading assessment text followed by a series of items, all of which relate to the

text. These types of units are common in PIRLS. Units are also regularly used in PISA. Depending on the assessment purpose, the units can be of different lengths and have varying numbers of items; however, what defines a unit is a set of items with a common stimulus. As with the multiple-choice items, the same stimulus followed by the same items, but in a different order, should be treated as different units, unless there is sufficient evidence to indicate that the order of the items does not affect the overall performance on the unit.

- *Blocks or clusters:* A block or cluster is a set of items, units, or some combination of the two, that is presented to examinees. In its simplest form, a block could contain just one item or one unit. Items (or units) are generally grouped into blocks to facilitate the assignment of items across different forms and to easily control content representation of the areas assessed across the forms. Grouping items into blocks helps maintain the context in which sets of items occur. The item context is a nuisance variable that has been shown to affect examinee performance. As an example of how organizing items into blocks helps, let us assume that we want to administer 10 items from a pool of 1,000 mathematics items—all multiple-choice with five response options—and let us assume that, of the 10 items we want, two items come from one of five mathematics subdomains. If we randomly pick 10 items from the entire pool, it is possible, although unlikely, that we could pick 10 items from only one or two of the subdomains, or 10 items with the correct answer being "A" or "B." Instead, we could organize the 1,000 items into 200 blocks of five items each. Here, each block would have one item from each of the subdomains, and each of the five response options would be used. Of course, we would not advertise that each of the five response options is used in every block, and this particular constraint might be difficult to implement. However, the test developer should certainly try to avoid undesirable and detectable patterns.

  When assembling blocks of items, the order in which the items are placed within the blocks could have an effect on the difficulty of the items. Although commonly used scaling models assume local independence of the items, local dependencies may be found. Verifying that the local independence assumption is not violated, as well the absence of position effects within or across blocks, is crucial. Analytic procedures rely on the assumption that the same item presented in two or more different positions can be treated as statistically equivalent instances of the same item.

- *Forms:* A test form is the actual set of items, in a specific sequence, that is administered to examinees. These items are organized in blocks, so technically a form is a set of blocks organized in a particular sequence. The same items administered in two different sequences are considered to be different forms. The term booklet is used to refer to the paper-and-pencil version of a form, whereas the term *form* covers paper-and-pencil administration as well as computer-based administration. The key consideration is that a different set of items or units, or the same set of items or units in a different sequence, is treated as a different form of the assessment and that evidence is gathered to show that the results on these items in the different forms are comparable.

- *Sessions:* The session is the particular time period during which a form, or part of it, is administered. Depending on its length, a form can be administered in one or more sessions. Most, but not all, large-scale assessments administer booklets in two sequential but distinct sessions. Each of the sessions is separately timed, and examinees are assigned a specific section of the form during each session.
- *Assessments:* The assessment constitutes the entirety of the item pool that is administered. Depending on the item pool and the purpose and duration of the program, the assessment could span multiple blocks, and multiple forms across administration cycles. Consider, for example, the PISA 2003 mathematics assessment. In this case, the total item pool consists of those items and units administered during the 2003 application. The PISA mathematics assessment item pool, however, would include all of the items and units administered across all four assessment cycles (2000, 2003, 2006, and 2009).

The concepts above can be represented as follows:

$(\text{Assessment}_a\ (\text{Form}_f\ (\text{Block}_b\ (\text{Unit}_u\ (\text{Item}_i)))))$.

In the case of some large-scale assessments, where the same assessment is administered in different languages, we would have the following:

$(\text{Assessment}_a\ (\text{Language}_l\ (\text{Form}_f\ (\text{Block}_b\ (\text{Unit}_u\ (\text{Item}_i))))))$.

$Item_i$ is thus nested within $Unit_u$, which is nested within $Block_b$, and so on. In its simplest form, an assessment can be constructed with just one form composed of one block that consists of one unit comprised of one item—a one-item assessment.

Booklet designs

The booklet design is the set of rules that we use to assign items to respondents. The set of rules could be as simple as assigning to every respondent all the items in the assessment in the same format and sequence. It is also possible to construct a sophisticated design that allows one to decide, during administration of the assessment, which item to administer next. This is the case in computer adaptive testing (CAT). Paper-based booklet designs require that these governing rules be defined prior to printing the forms. Computer-based assessments are more flexible with respect to design choices because algorithms can be used to decide—even as the assessment is being administered—which specific items to administer next.

An example

Most large-scale educational assessments (e.g., TIMSS, PIRLS, and NAEP) measure a defined set of skills within a representative sample of the population of interest. The goal of these assessments is to describe groups within the population with respect to broadly defined areas of school- or work-relevant skills. A commonality of these assessments is that individual scores are not assigned to examinees. Although this feature might be seen as a limitation, it actually allows for more flexibility in the choice of design.

Take the example of TIMSS 2007, which included in its Grade 8 mathematics assessment content domain number, *algebra*, *geometry*, and *data and chance*. Across these wide content domains, examinees were expected to draw on the cognitive domains of *knowing*, *applying*, and *reasoning* (Mullis et al., 2005). This approach brought the total number of reporting scales to eight (seven overlapping subscales plus overall mathematics). Reporting in this fashion implies that an ample number of items in each of the reporting sub-domains is administered across the population, such that sufficiently precise estimates of proficiency distributions are possible.

In total, the TIMSS 2007 mathematics assessment consisted of 215 items, a number that made it impossible to test all examinees on every item. (Estimates for completing one item ranged from one minute for the multiple-choice items, to three to four minutes for the open-ended items.) Instead, the 215 items were distributed across 14 mathematics blocks; each examinee received two of these. This design ensured linking across booklets because each block (and therefore each item) appeared in two different booklets. Booklets were then administered to randomly equivalent samples of examinees, such that the total assessment material was divided into more reasonable 90-minute periods of testing time for each examinee. With this design, examinees received just a small subset of the total available items.[2] (For complete details on the TIMSS 2007 booklet design, see Olson, Martin, & Mullis, 2008.)

While a multiple matrix booklet design ensures coverage of a broad content domain in a reasonable amount of time, it poses challenges associated with putting the items onto a common scale, estimating examinee proficiency, and ultimately obtaining population estimates. Advanced statistical techniques are available to estimate the distribution of proficiencies in populations and subpopulations of examinees. Mislevy (1991), Mislevy, Beaton, Kaplan, and Sheehan (1992), and Mislevy, Johnson, and Muraki (1992) describe these methods. Description and discussion of more recent developments can be found in von Davier, Sinharay, Oranje, and Beaton (2006) as well as in Adams and Wu (2007).

Selecting a booklet design

Selecting a suitable booklet design for the reporting needs of the assessment program requires careful consideration of many factors and an analysis of the advantages and disadvantages of the different options that are available and possible. The exact booklet design is a response to the specific needs of the assessment program, and a booklet design that is suitable for one program might not be suitable for another. Note that we qualify the design chosen as "suitable" rather than as "right" or "correct." We do this because the booklet design will only be "right" or "correct" to the extent that it is able to address all the needs of the assessment program. As we will see, the "right" booklet design will always be a result of numerous compromises based on a multitude of factors, some of which we mention below.

2 We note here that when blocks are rotated across the different booklets, an assumption is made that the examinee will have time to reach all the items in the book, and that fatigue or other effects that may affect performance will not have set in.

The guiding factor in developing and choosing a booklet design is the purpose of the assessment. A critical question that needs to be asked during the development process is if the assessment is intended for selection, assigning grades, diagnosis, or for describing a population? When the assessment is designed for the purpose of individual or group decision-making, more measurement precision is necessary, which means that more items generally need to be included in the particular forms administered. In general, the booklet design should allow for the administration of a sufficient number of items on the domain of interest to ensure that the desired statistical precision requirement is met; however, in this regard, there are no absolutes. If the purpose of the assessments is to make decisions about individuals, as in admission tests, it is critical to ensure that individuals are measured with a high level of precision. This precision should, to the extent possible, be relatively uniform across the individuals, particularly around the points in the proficiency continuum where decisions will be made. However, if the purpose of an assessment is to describe skills of subgroups of interest, less precision is necessary at an individual level; sufficient precision must nevertheless be achieved at the subgroup level.

Once the purpose of the assessment is decided, a natural next choice is to consider how broad a domain we want to measure. For example, a Grade 8 student's assessment could include only content covered during Grade 8 or content covered up to and including material covered in Grade 8. A third option could include an assessment that measures content necessary to advance to Grade 9. Deciding on the assessment breadth also gives an idea of the breadth of the content that would need to be included in the assessment. In addition, it is important to consider the number of reporting domains and subdomains. For example, it may be desirable to report only overall mathematics achievement or, additionally, achievement in algebra, geometry, and numbers. The types of knowledge and skills assessed also help one make decisions about how many items are necessary to reliably measure the domains of interest. For example, reading assessments at lower grades usually consist of a one- to three-page text followed by a set of items referring to the text. Because of the time needed to read the text, it is not possible to administer, to any one individual, the numbers of items that can typically be administered during a mathematics or science test. The nature of a writing assessment usually limits the number of writing tasks to two or three during one session. Depending on the area assessed, the booklet design might accommodate more or fewer items.

The number of reporting scales is another important consideration in the booklet design. In general, a positive relationship exists between the number of reporting scales and the number of items that need to be included in the assessment (Embretson & Reise, 2000). As a general rule, and depending on the characteristics of the test items (mainly discrimination and difficulty), it is advisable to include 20 to 30 items per subscale in any given administration of a survey assessment (von Davier, Gonzalez, & Mislevy, 2009). But this choice ultimately depends on the content that needs to be covered as part of the domain. The content coverage requirement and measurement precision will provide the general guidelines for determining the optimal number of items per domain.

The topics in the domain of interest should be proportionately represented in the assessment. However, the number of items necessary is inversely related to their quality, as measured by the amount of information they provide. Nonetheless, it is important to consider that adding low discriminating items or off-target items does not make for better measurement, and that too few items, even if on target and discriminating, are unlikely to capture, in a useful way, the breadth of the domain measured. When there are five subscales, for example, using a rule of 20 to 30 items per subscale leads to an assessment of about 100 to 150 items. Clearly, these numbers represent too many items to administer to any single student within a reasonable period of time, assuming we want to avoid a speeded assessment.

Once we have determined how many items are necessary to cover the domain of interest, our next considerations are the time available to administer the tests, and the time necessary to complete the tasks in the assessment. We need to know not only how much time is available but how many items can reasonably be administered and answered during a given time period. Piloting the items will provide a good estimate of how much time is needed to answer the questions in the assessment. Items with a heavy reading load generally take longer, and younger examinees generally take longer because of the reading involved. Older examinees tend to be faster readers, and can also maintain their focus and attention for longer periods of time. Consequently, it is reasonable to test adults for longer periods.

Among the other matters that need to be considered when determining a suitable booklet design is how much time is available to administer the tests. In international studies such as TIMSS and PIRLS, participation in large-scale assessments does not carry high-stakes consequences for examinees or schools, and participation is voluntary. An overly time-consuming assessment will only discourage participation, and will be seen more as a disruption to the school day rather than as an opportunity to learn what examinees know and can do. Another time consideration is the length of the school's class and recess periods, which often need to be observed to minimize disruptions and distractions during test administration. Many schools around the world organize their day in 45- to 60-minute periods, with longer periods necessitating a break at some point. It is because of period lengths that assessments such as NAEP use 50 minutes of testing time. Other assessments, such as PIRLS, administer the booklets for Grade 4 in two 40-minute sessions with a 5- to 10-minute break in between. TIMSS uses two 45-minute sessions for Grade 8, with one 5- to 10-minute break in between. PISA, administered to 15-year-olds, uses two 60-minute sessions, with a 5- to 10-minute break in between.

It is also important to consider, with respect to the number of reporting scales in the assessment, how the actual results will be reported. For instance, will a unidimensional skill variable be reported in the assessment or will there be multiple proficiency variables, one for each of the subscales? As a general rule, the greater the number of variables that are reported, the greater the number of items needed to reliably estimate these scores across the population.

An additional and related consideration is whether results are reported as simple scale scores or as attainment of educational standards in the form of cut-scores or benchmarks in the achievement distribution. It is important, when reporting whether examinees have reached a particular point in the distribution, to make sure that the number and the quality of items administered across the population allow us to make such inferences with respect to the domain. A form generated dynamically and based on examinee responses to the items administered initially will allow us to effectively tailor an assessment to the level at which the examinee is ultimately classified, provided appropriate content-control algorithms are used. However, a static form administered to an examinee will need to measure, with an acceptable level of precision, different points in the distribution to make it possible to infer whether or not the examinee has met or surpassed such points in the distribution.

One last consideration is the number of times that the assessment will be administered. Assessments that are administered only once often require a simpler booklet design, whereas ongoing cyclical assessment programs require careful consideration of issues related to item-release policies, linking scores from one administration to the next and across all administrations, renewal of the framework, refreshment of the item pool, introduction of new components in the assessment, and last, but not least, risk of disclosure of the items and test security. The greater the number of administrations planned for an assessment, the greater the need for carefully considering all the issues mentioned above so that proper linking of the results from the administrations can take place, and so that adequate inference can be made from the assessment results.

Deciding on a booklet design depends on the answer to the questions posed above, and perhaps to some others that are unique to the particular assessment programs. Except for simple cases, no single booklet design will fit all programs, or will fit any one program. The design is always a compromise between what is desired and what is possible, given the specific circumstances and resources available. In a recent article, Frey, Hartig, and Rupp (2009) discuss in more detail many of the constraints that need to be considered when choosing a design.

Examples of booklet designs

Although booklet designs are specific to the assessment, we can classify them for explanatory purposes, and to elucidate a number of advantages and disadvantages of each kind.

In general, we can classify booklet designs into two overarching categories—*complete* and *incomplete* designs. In *complete* matrix booklet designs, all the forms contain all the items, and therefore all examinees are administered all items in the assessment. *Incomplete* matrix booklet designs are characterized by a design in which any one form has only a subset of the items, and any one examinee is administered only a subset of the total number of items. Even though a complete matrix booklet design requires all examinees to take all the items, the existence of multiple forms can rotate the position of the blocks within the forms. Reasons for this practice include

controlling for position effects within the booklet or form, and preventing examinees from copying one another's responses during test administration.

A *balanced* booklet design is one in which each block of items is rotated to appear an equal number of times in each position within the forms across the entire booklet design. This feature is a desirable one if an order effect is suspected (i.e., the order of the items has an influence on examinees' responses to the items). A *balanced* booklet design controls for such an effect.

Tables 2 and 3 show examples of complete and incomplete matrix booklet designs, respectively, for a three-block assessment. In each table, the top set of rows shows an unbalanced booklet design, whereas the bottom set of rows shows a balanced booklet design. Under the incomplete booklet design, each form consists of only two blocks, or two-thirds of the assessment. This allows for a one-third reduction in test time or for the inclusion of 50% more items in the same amount of time as the complete booklet design. However, note that the first incomplete booklet design shown in Table 3 does not include a form in which Blocks A and C are administered together. As a consequence, we cannot directly compute the correlation between the items from these two blocks—an important aspect that many multivariate methods rely on and something the balanced booklet design shown will allow. An added complication might arise if Block B occurs in all forms. Because Blocks A and C occur in only one of the two forms, the precision of the parameter estimates for A and C will be less than that for Block B.

Table 2: Example of complete matrix booklet designs

Form	Blocks		
1	A	B	C
Form	Blocks		
2	A	B	C
3	B	C	A
4	C	A	B

Table 3: Example of incomplete matrix booklet designs

Form	Blocks	
1	A	B
2	B	C
Form	Blocks	
3	A	B
4	B	C
5	C	A

Incomplete booklet designs can also be generated dynamically, as is the case with some computer-delivered tests. Table 4 shows three booklet designs, each with different numbers of blocks. In this example, an examinee would be first administered an "average" block and, based on his or her performance on this block, would then be presented with an easier or more difficult block. Imagine a very able examinee who excels at an average block. The next block administered would be a "difficult" block. If the examinee performed poorly on this block, a somewhat less difficult block would be administered.

Table 4: Example of dynamic incomplete matrix booklet designs

Very easy	Easy	Somewhat easy	Average	Somewhat difficult	Difficult	Very difficult
A	B	C	D	E	F	G
A B	C D	E F	G H	I J	K L	M N
A B C	D E F	G H I	J K L	M N O	P Q R	S T U

When such an assessment is implemented, it is important that the screening block is of very high reliability to ensure the examinee is not routed to blocks that are too easy or too difficult. Including several blocks at each level also allows booklet designers to broaden the coverage of the domain. This practice minimizes exposure of items and prevents subsequent examinees from being administered the same sequence of items during an administration, thus lessening opportunities for cheating.

One booklet design that has many desirable characteristics is the 7-block Youden squares design, originally used in experimental biological research designs (Preece, 1990). This design is often referred to as the BIB7 design. Table 5 shows this design, which was first used in an educational assessment context in NAEP. Extensions of this design can also be found in assessments such as TIMSS and PISA. The design consists of seven blocks and seven forms, with each form containing three blocks. Each block appears once in each position in the design, and each block also appears once with each of the other blocks. Despite the desirable characteristics of the BIB7 design, the fact that there is an odd number of blocks for each form makes it unsuitable for administration in assessment situations involving two sessions with a break in between. Depending on the number of items within the blocks, and the time needed to answer them, this design is better suited for assessments conducted over one or three sessions.

Table 5: Balanced incomplete 7-block design (BIB7 or Youden squares design)

Form	Block		
1	A	B	D
2	B	C	E
3	C	D	F
4	D	E	G
5	E	F	A
6	F	G	B
7	G	A	C

The designs shown above illustrate just some of the possibilities available, and many of them can and often need to be extended over time. Table 6 shows a BIB7 design extended over four years of an annually administered assessment. In this design, one of the forms is released to the public every year. The released blocks are then replaced the next year. In the table, the released blocks are underlined, and the new blocks appear in bold. Note that as the assessment progresses from year to year, the item pool is renewed with new blocks and items. Year 4 of the assessment includes one block from Year 1 (G1), one from Year 2 (A2), two from Year 3 (B3 and E3), and three blocks from Year 4 (C4, D4, and F4). After the release of Form 4 in Year 4, the Year 1 items will no longer appear in the design.

Table 6: Balanced incomplete 7-block design over four years

Form	Year 1			Year 2			Year 3			Year 4		
1	*A1*	*B1*	*D1*	**A2**	**B2**	**D2**	A2	**B3**	D2	A2	B3	**D4**
2	B1	C1	E1	***B2***	C1	E1	**B3**	**C3**	**E3**	B3	**C4**	E3
3	C1	D1	F1	C1	**D2**	F1	***C3***	*D2*	*F1*	**C4**	**D4**	**F4**
4	D1	E1	G1	**D2**	E1	G1	D2	**E3**	G1	***D4***	*E3*	*G1*
5	E1	F1	A1	E1	F1	**A2**	**E3**	F1	A2	E3	**F4**	A2
6	F1	G1	B1	F1	G1	**B2**	F1	G1	**B3**	**F4**	G1	B3
7	G1	A1	C1	G1	**A2**	C1	G1	A2	**C3**	G1	A2	**C4**

AN EXAMINATION OF PARAMETER RECOVERY USING DIFFERENT BOOKLET DESIGNS

Analytical Approach

We used the above examples of different booklet designs to examine how these designs influence the recovery of the parameters that were used to generate the data. In particular, we used a simulated dataset to investigate how well we could recover the generating parameters for the population and the items, and to determine the effect of these designs on reporting group-level results. Knowing the generating values is useful when comparing estimates that try to recapture these values.

For our analysis, we generated mathematics proficiency-skill levels for 4,000 cases crossed with two known background characteristics—school type with Levels A and B, and socioeconomic status (SES), also with two levels, high (H) and low (L). This approach resulted in four (2x2) distinct groups, each with 1,000 cases. We simulated the average difference in mathematics skills between School Types A and B to be 0.000, and we set the average difference based on parental SES to 1.414. The average ability level was +0.707 for the SES H group, and -0.707 for the SES L group. Let us for now ignore considerations as to whether these assumptions about school and SES differences are particularly realistic or unrealistic, especially given that these variables may show different effects in different populations.

Table 7 presents the means and standard deviations that we used to generate the response data. We set the standard deviation within each of these groups to 0.707, which yielded a variance within each of the four groups of about 0.5 (or 0.707^2), and an overall variance and standard deviation of 1.000. The data that we used in this analysis were the same as those that von Davier et al. (2009) used in their analysis.

Table 7: Means and standard deviations (in parenthesis) used to generate the simulated dataset

		School		
		A	*B*	*Total*
SES	L	-0.707 (0.707)	-0.707 (0.707)	-0.707 (0.707)
	H	+0.707 (0.707)	+0.707 (0.707)	+0.707 (0.707)
	Total	0.000 (1.000)	0.000 (1.000)	0.000 (1.000)

We simulated the responses of all examinees to a pool of 56 items, assuming a two-parameter logistic response model (2-PL). The 2-PL describes the probability of a correct response to an item as a function of the person's ability, and the discrimination and difficulty parameter of the item. In the 2-PL IRT model, the probability of a correct response is given by:

$$P(x_i=1 \mid \theta_k, a_i, b_i) = \frac{1}{1+\exp^{(-1.7\, a_i(\theta_k - b_i))}}$$

where

x_i is the response to item *i*, and 1 if correct and 0 if incorrect;

θ_k is the proficiency of an examinee on scale *k*;

a_i is the slope parameter or discrimination of item *i*; and

b_i is the location parameter or difficulty of item *i*.

Item difficulties were distributed uniformly between -1.0 and +1.0. Item discriminations ranged between 0.5 and 1.5. No correlation was built in between difficulty and discrimination, and the items were randomly assigned to each block.

We simulated the responses to the items under four different conditions:

1. All examinees were administered the 56 items.
2. Items were randomly assigned to one of seven blocks, labeled A, B, C, D, E, F, and G. Based on the form booklet design, every examinee was administered three blocks (24 items in total) in the assessment pool. The design organized the blocks according to the BIB7 design described earlier. The resulting forms were (ABD), (BCE), (CDF), (DEG), (EFA), (FGB), and (GAC).
3. Every examinee responded to two blocks (16 items in total) in the assessment pool; the blocks used were the same as those composed for (2). The blocks were organized into seven pairs as follows: (AB), (BC), (CD), (DE), (EF), (FG), and (GA).
4. Items were randomly assigned to one of 14 blocks, named A through N. Every examinee responded to two of these blocks (eight items in total) in the assessment

pool. These blocks were organized into 14 pairs as follows: (AB), (BC), (CD), (DE), (EF), (FG), (GH), (HI), (IJ), (JK), (KL), (LM), (MN), and (NA).

No block order effects were introduced into the simulation. The item parameters used to generate the data thus stayed the same for the items, even when administered in different block positions.

We next calibrated the items with PARSCALE Version 4.1 (Muraki & Bock, 1997), using marginal maximum likelihood estimation procedures. We used EAP (expected a posteriori) estimators to compute the reliability estimates for each examinee, and we then compared the results for each of the four conditions in terms of their ability to recover the generating parameters. Because EAP scores are point estimates, aggregating EAP scores leads to underestimation of the variance of the population. We compared achievement means and standard deviations overall and by subgroups. We also computed correlations between the estimates and the generating parameters, and constructed plots to display graphically the differences and the effects of the different designs.

Results

We present the results in two sections. In the first, we present the comparison between the generating ability and the EAP estimates obtained from the analysis under each of the four conditions. In the second section, we present the comparisons between the generating item parameters and the estimates of the item parameters. In total, we conducted 100 simulations. The results presented in the tables that follow are the average results for the simulations.

Comparing person ability estimates

Table 8 presents the number of cases within each simulation group, defined by the number of items administered to the examinees, the average true and estimated person ability of these cases, and their corresponding standard deviations. Tables 9 and 10 present the same results, but broken down by the two background variables used to generate the simulated data. We note again that because we simulated no school effect, the overall means by school are the same. We did, however, simulate a SES effect, such that those in the group SES Type H were more able than those in SES Type L.

Table 8: Means and standard deviations overall

Number of items	*N*	Average true score	Average EAP score	Standard deviation of true ability	Standard deviation of EAP estimate
8 items	4,000	-0.010	0.007	0.994	0.872
16 items	4,000	-0.010	0.010	0.994	0.933
24 items	4,000	-0.010	0.012	0.994	0.956
56 items	4,000	-0.010	0.015	0.994	0.984

Table 9: Means and standard deviations, by school

Number of items	School	*N*	Average true score	Average EAP score	Standard deviation of true ability	Standard deviation of EAP estimate
8 items	A	2,000	-0.027	-0.006	0.980	0.868
16 items	A	2,000	-0.027	-0.004	0.980	0.925
24 items	A	2,000	-0.027	-0.002	0.980	0.947
56 items	A	2,000	-0.027	-0.002	0.980	0.972
8 items	B	2,000	0.007	0.020	1.008	0.876
16 items	B	2,000	0.007	0.024	1.008	0.940
24 items	B	2,000	0.007	0.025	1.008	0.965
56 items	B	2,000	0.007	0.031	1.008	0.995

Table 10: Means and standard deviations, by SES

Number of items	SES	*N*	Average true score	Average EAP score	Standard deviation of true ability	Standard deviation of EAP estimate
8 items	H	2,000	0.686	0.559	0.721	0.674
16 items	H	2,000	0.686	0.632	0.721	0.669
24 items	H	2,000	0.686	0.659	0.721	0.708
56 items	H	2,000	0.686	0.694	0.721	0.720
8 items	L	2,000	-0.706	-0.545	0.698	0.676
16 items	L	2,000	-0.706	-0.611	0.698	0.691
24 items	L	2,000	-0.706	-0.636	0.698	0.699
56 items	L	2,000	-0.706	-0.664	0.698	0.703

In general, Tables 9 and 10 show that the EAP estimate of the mean overall and by school, where there were no simulated group differences, matched the overall true mean of the simulated data. However, one noticeable difference between the EAP scores and the generating abilities was that as the number of items decreased, so too did the variability of the estimated posterior means of abilities. The latter reached only 0.87 when eight items were administered, whereas the standard deviation for the generating abilities was 0.994, overall. We observed similar differences when we broke down the results by school type. When we inspected the results by SES, we noticed not only that the variability decreased, as it did overall and by school type, but also that the means for these two groups and, as a consequence, the differences between these two groups, diminished. The reason behind this effect is explained in von Davier et al. (2009).

In Table 11, we present the student-level correlation between the generated simulated scores and the ability estimates. Consistent with the previous results, the reduction in variation as the number of items administered decreased paralleled a decrease in the correlation between these variables. (These values are also a measure of the reliability of the measurement.)

Table 11: Correlations between "true" and estimated ability

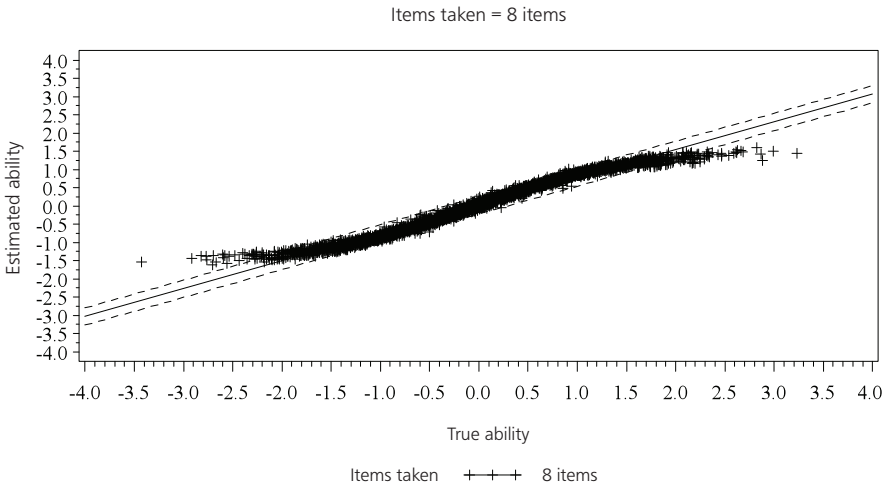
Number of items	Correlation
8 items	0.869
16 items	0.925
24 items	0.946
56 items	0.974

Figure 1 presents the plots of the ability estimates obtained from the simulations under the two extreme conditions (8 items and 56 items per person), plotted against the true ability used to generate the data. The plots include the regression line and 95%-confidence intervals for the data. These plots provide further evidence that as the number of items administered increased, so too did the correspondence between the estimated ability and the true ability of the subjects. Figure 1 allows us to observe graphically what the previous tables illustrated numerically: as the number of items increased, the variability of the scores also increased, approaching the true variability of ability, while the distribution of true scores at any one point on the continuum of the estimated abilities diminished, indicating better precision of the estimates.

Figure 2 shows the plots of the standard errors of the EAP scores, again from one simulation, plotted against the EAP estimate, under the two extreme conditions. As we expected, the estimates toward the extreme of the distribution, where measurement was *less* precise, had larger standard errors, while the estimates toward the center of the distribution, where measurement was *more* precise, tended to be smaller. The higher precision toward the center of the distribution occurred because the difficulties of the item parameters used in this simulation ranged between -1.0 and +1.0, resulting in more precise measurement in this area of the distribution. Also worth noting in these plots is the pattern that we expected: as the number of items increased, the error of the estimates decreased. But notice also that the decrease or increase in the average error is not proportional to the change in the number of items. For example, the average error at the center of the distribution ranges from 0.40 to 0.50 when 8 items are administered, whereas it is about 0.18 when all 56 items are administered.

The last set of results comparing person ability estimates, perhaps the set that should raise more concerns among those interested in estimating group differences using matrix booklet designs, is shown in Figures 3 through 6. Again, remember that in our simulation we introduced a SES effect, but there were no differences between school types. In these figures, we plotted the true and estimated average difference between examinees from each of these groups (School A and B or SES types L and H), at each percent-correct score (of the total items) in the distribution. (We use the percent-correct metric in these plots because it better illustrates our point.) Note that in Figures 3 and 4 there are virtually no differences by school type in the theta metric, except in the extremes. These differences were eliminated in the EAP metric. Thus, examinees at any percent-correct point of performance along the distribution received about the same estimated score regardless of the type of school they were attending (Figure 4); this is what we would expect, given the true scores (Figure 3).

Figure 1: Plots of "true" ability against estimated ability, by numbers of items administered



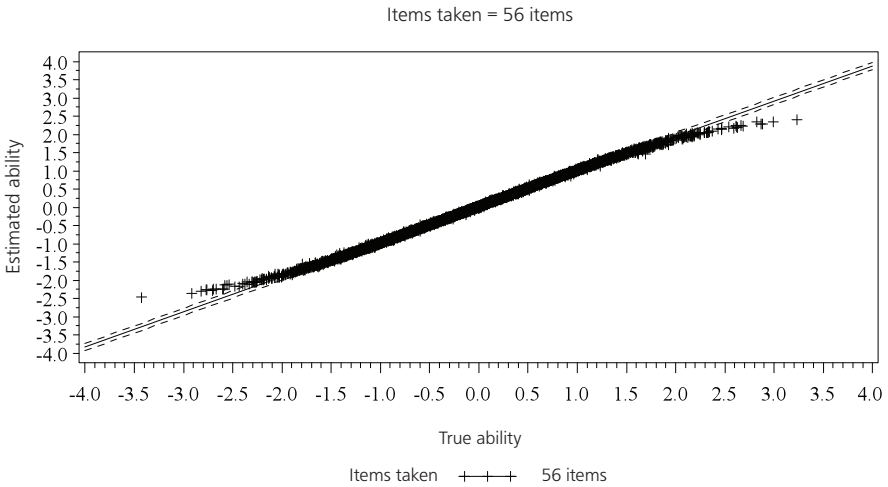


Figure 2: Plots of estimated ability against the standard error of the estimate, by numbers of items administered

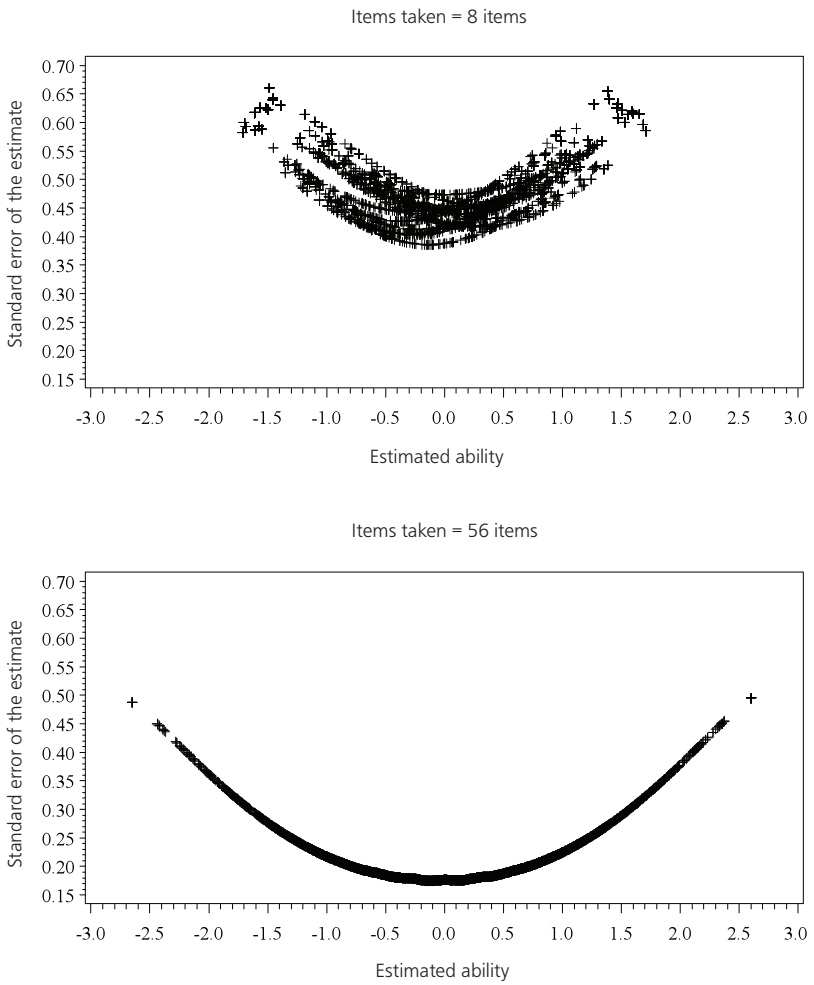


Figure 3: Plot of true (theta) average differences between examinees from Schools A and B, by percent correct

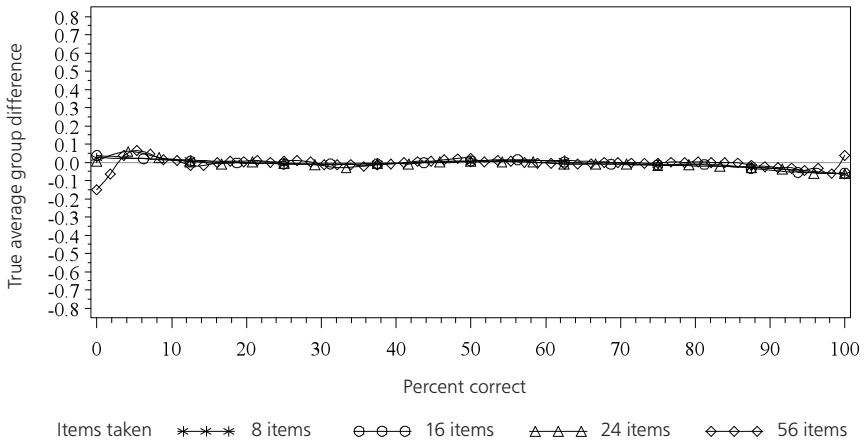
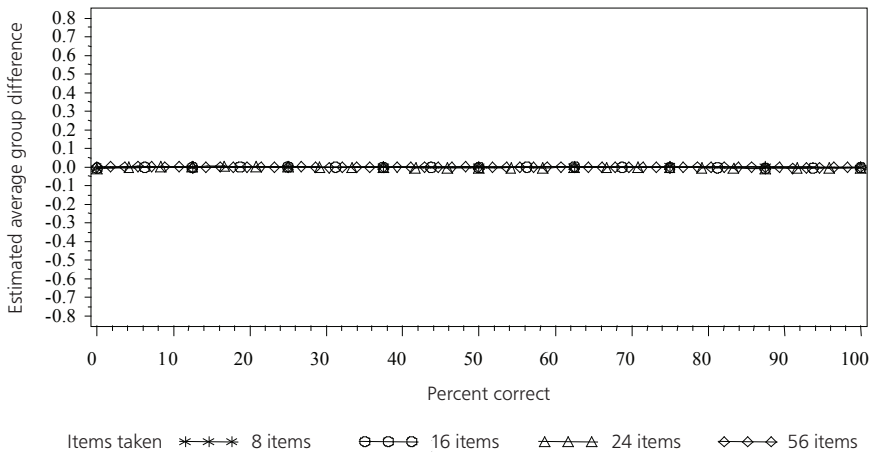


Figure 4: Plot of estimated (EAP) average differences between examinees from Schools A and B, by percent correct



However, when we look at these results by SES type in Figures 5 and 6, we notice that although examinees who obtained the same percent-correct score (as is clear from the true scores in Figure 6) had, on average, different true abilities, depending on their SES type, these differences did not show up in the EAP scores. In other words, real group differences were underestimated, and the degree to which these differences were underestimated increased noticeably as the number of items administered decreased.

How one can properly estimate these differences is beyond the scope of this paper, but it is the subject of the paper by von Davier et al. (2009). Of concern to us in this paper is the fact that, as the number of items decreased, our estimates of a person's ability became less reliable, and we underestimated group differences in the population, when these existed. These findings likely lead the reader to ask, "How many items are necessary?" The answer is, "It depends"—on the number of items needed to sufficiently cover the domain of interest and on the quality of the items. We can only say for certain that there was a meaningful relationship between the number of items and estimate precision. For individual score reporting, the 24 and 56 items seemed to yield sufficient reliability. Note that even in the 16-item case, we estimated a reliability of above .9. However, in real applications, this estimate might not be as easy to achieve. Simulated data are far cleaner in the sense that these simulated examinees produce responses that follow the model perfectly.

Comparing item parameter estimates

When reviewing the results of the item parameter estimates, keep in mind that because of the rotation of each block within each design, the items were administered to samples of varying sizes. When all 56 items were administered to all examinees, each item was attempted by the total sample of 4,000 people. With a 24-item administration, each block appeared in three of seven total books. As such, each item had about 1,700 responses under this design. In the designs with 16 and 8 items, each item had just 1,150 and 570 examinee responses per item, respectively. To maintain consistency with the previous section, we present the results for the item parameter recovery by number of items administered to an individual.

Tables 12 and 13 present descriptive statistics for the difficulty and discrimination parameters. While the estimates did not exactly match the true parameters, we found no particular pattern in the data. In fact, we can see from the tables that even when the examinee sample size was reduced to 570, as was the case when each individual was administered only eight items, the distribution of the estimated item difficulties and discrimination was, on average, fairly close to the true difficulty and discriminations. Table 14 provides further evidence of this finding. Here, we can observe that the correlation between the estimated parameters and the true parameters, particularly the difficulty parameters, was close to 1.0. This correspondence was also true of the discrimination parameter; however, the correlation decreased slightly as the number of respondents decreased.

Figure 5: Plot of true (theta) average differences between examinees from SES L and H, by percent correct

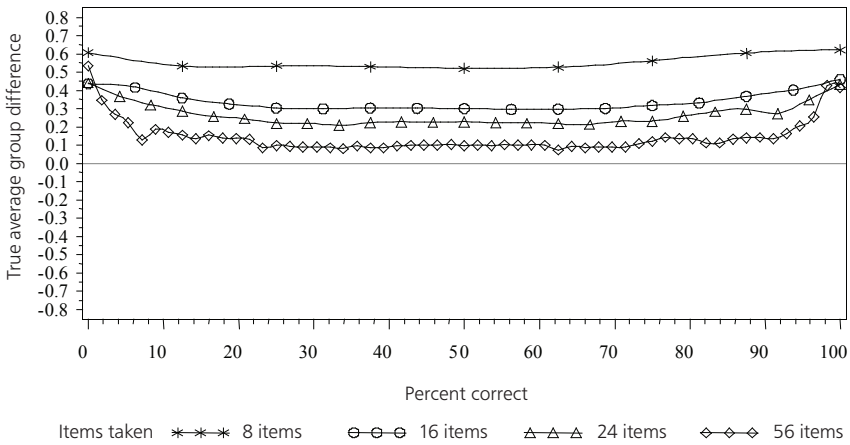


Figure 6: Plot of estimated (EAP) average differences between examinees from SES L and H, by percent correct

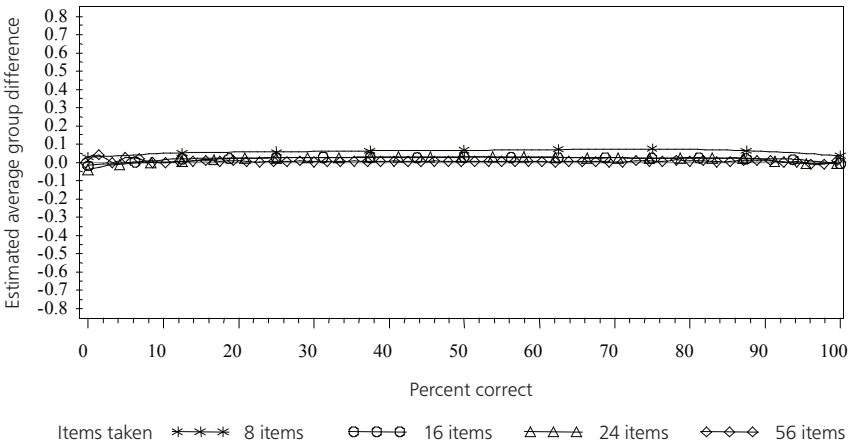


Table 12: Descriptive statistics of item difficulty parameter estimates, by number of items administered

Number of items	Average true difficulty	Average estimated difficulty	Standard deviation of true difficulty	Standard deviation of estimated difficulty
8 items	-0.064	-0.032	0.529	0.531
16 items	-0.064	-0.024	0.529	0.549
24 items	-0.064	-0.034	0.529	0.516
56 items	-0.064	-0.040	0.529	0.536

Table 13: Descriptive statistics of item difficulty parameter estimates, by number of items administered

Number of items	Average true discrimination	Average estimated discrimination	Standard deviation of true discrimination	Standard deviation of estimated discrimination
8 items	0.929	0.951	0.249	0.260
16 items	0.929	0.927	0.249	0.264
24 items	0.929	0.939	0.249	0.266
56 items	0.929	0.931	0.249	0.253

Table 14: Correlations between item parameter estimates and true parameters

Number of items	Correlation of difficulty	Correlation of discrimination
8 items	0.983	0.936
16 items	0.994	0.966
24 items	0.996	0.972
56 items	0.999	0.992

Figures 7 and 8 support the previous findings with respect to the fairly close match between the difficulty estimates and the true difficulties, regardless of the condition. To save space, we present only the plots for when students were administered 8 and 56 items. The plots of conditions in between simply showed a continuation of the observed pattern.

Figure 7 presents the plots of the estimated item difficulties with the true item difficulties, along with the linear regression line and the 95% confidence interval. Here, and again as we expected, as the number of respondents to the items increased, so too did the precision with which the item difficulties were estimated. In Figure 8, which plots the difficulty estimates against the error of the estimates, we can observe the expected pattern of a slight U-shaped spread, indicating that the difficulty estimates for the more difficult and easier items were not as precise as the estimates of the items toward the middle of the continuum—the area containing most of the distribution of respondents. Nonetheless, as the number of people responding to the items increased, the precision increased, particularly at the extremes of the distribution.

Figure 9 shows the plots of the discrimination parameter estimates against the true discrimination parameters, and Figure 10 illustrates the plots of the discrimination parameter estimates against the error of the estimate. As with the difficulty parameter, we can see that the estimates were well matched to the distribution of true parameters, and that this match improved as the number of items administered to any one individual—and, as a consequence, the number of respondents per item—increased. A pattern different from that observed with the difficulty estimates is the positive and linear relationship between the item discrimination estimate and the error of the estimate. However, this relationship weakened as the number of respondents per item increased. This was because more information from responses was available for use in the estimation of the discrimination parameter.

Figure 7: Plots of difficulty estimates against "true" difficulty, by number of items administered

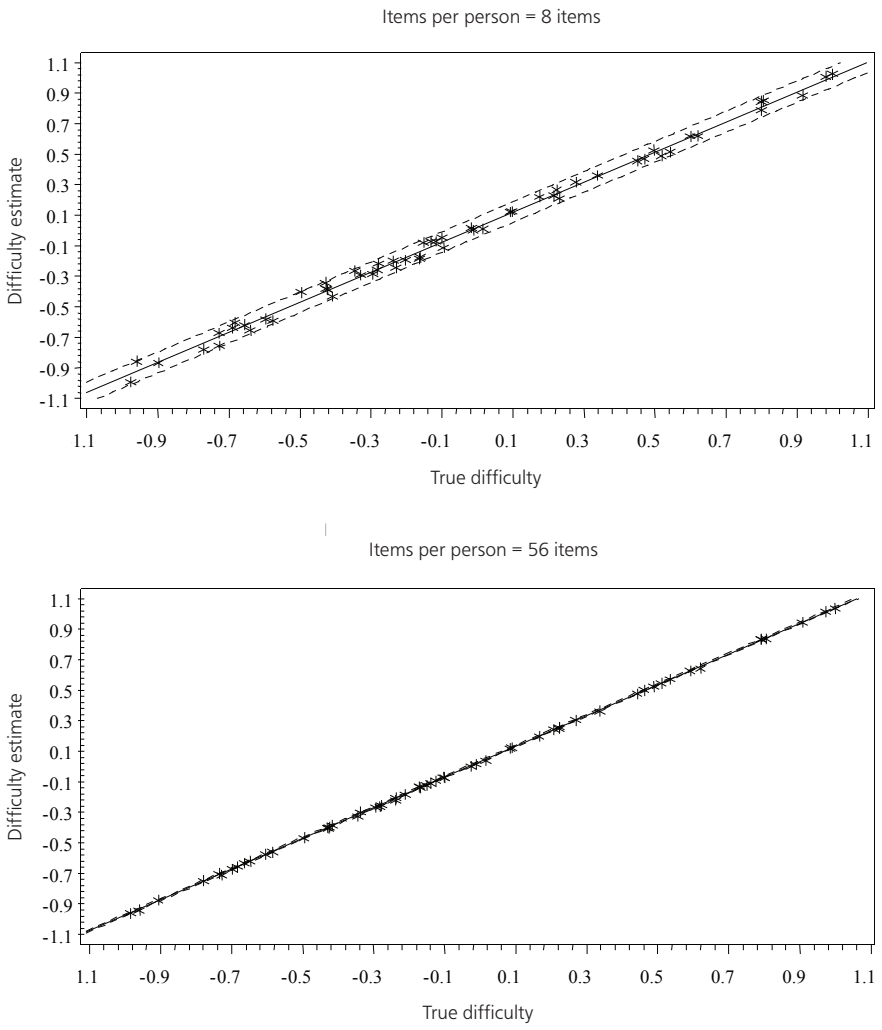


Figure 8: Plots of estimated difficulty against the standard error of the estimate, by number of items administered

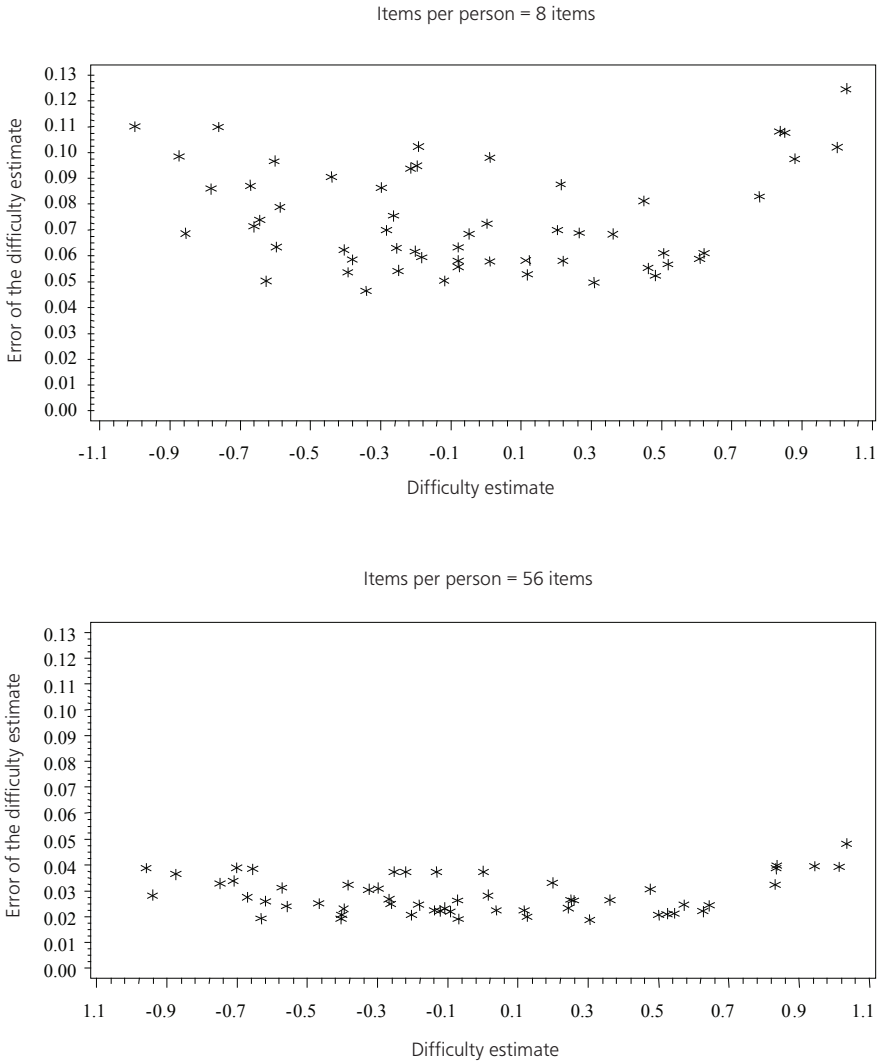


Figure 9: Plots of discrimination estimates against "true" discrimination, by number of items administered

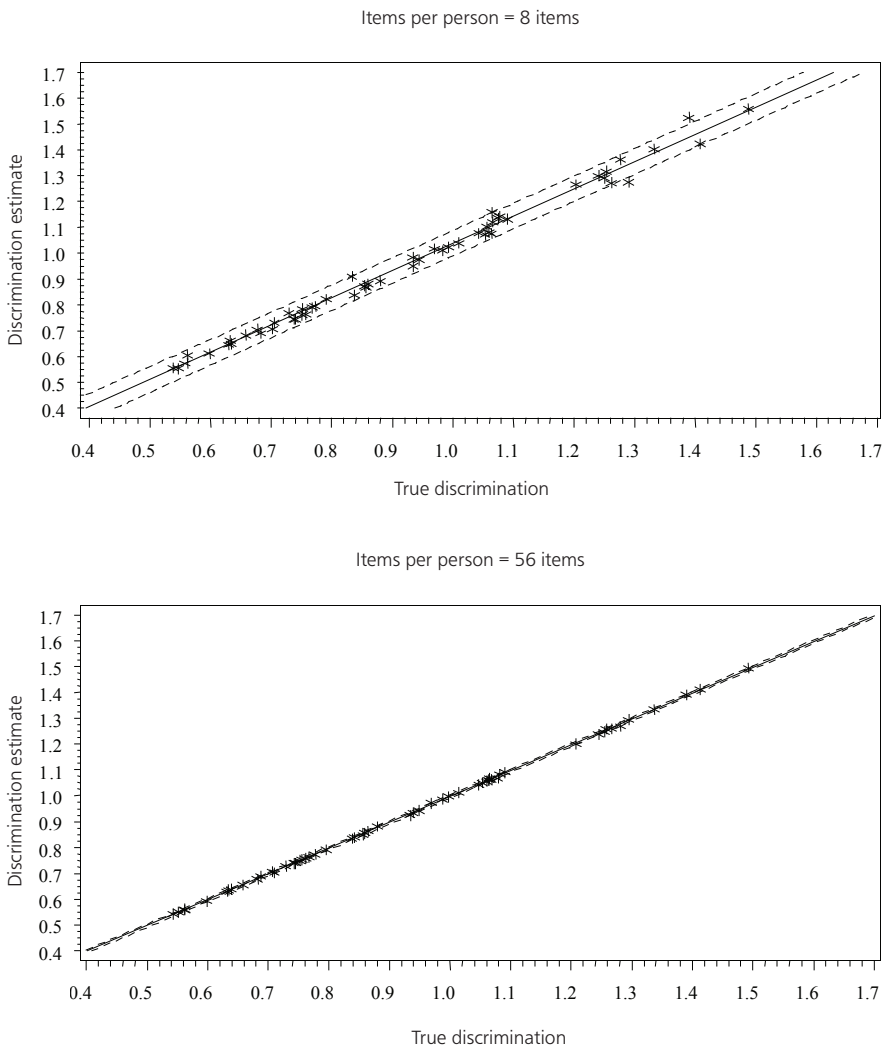
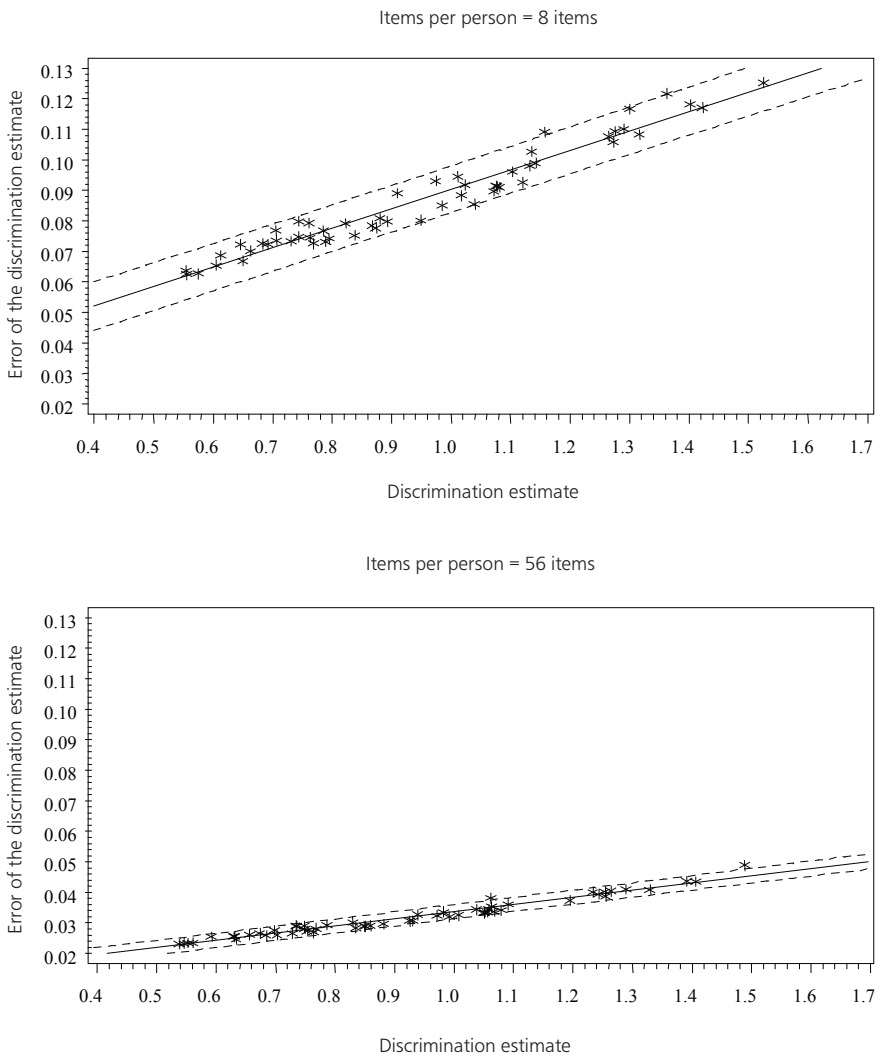


Figure 10: Plots of estimated discrimination against the standard error of the estimate, by numbers of items administered



CONCLUSIONS AND RECOMMENDATIONS

In this article, we have discussed issues that test developers and other interested parties need to consider when developing a booklet design. Underlying all of the issues that we have mentioned is the purpose and use that will be made of the assessment results, the breadth of the content that is intended to be covered, the time available for administering the assessment, and the precision of the resulting estimates that is expected or tolerable. We also presented a few examples of multiple matrix sampling booklet designs and discussed their advantages and disadvantages. Finally, we used simulated data to describe the effects of a few different simulated designs on the parameter estimates obtained.

In general, we hope that readers understand that choosing one design over another is a decision based on trade-offs. More precision is obtained with more items, but these do require more response time per examinee and more time and expense in terms of item development. Fewer items administered per person results in less precise estimates; however, respondent burden is reduced. In the end, test developers and psychometricians need to come together to explore the consequences of the different options and to develop a design that ensures proper coverage of the domain assessed and proper coverage of the population assessed. They also need, throughout this process, to ensure that the levels of precision achieved are acceptable with respect to the reporting purpose. Using simulated data or using item and examinee samples from real data collected in previous assessments will help answer some of these questions.

In discussing the results presented in this paper, we should point out that they are based on simulated data with specific characteristics, and therefore in some sense are relatively clean compared to what can be expected from real data stemming from operational data collection. It is likely, therefore, that different results would be obtained if real assessment data were used. For example, in the simulations, we used a well-targeted set of items, with difficulties well within the range of the abilities of the respondents. These items, moreover, were administered to random samples of respondents, with abilities normally distributed around the area where items provide more information.

We did not include, in our simulations, conditions such as order effects, skewed score distributions of students, and mismatch of the test information such as curve to the ability of the population of interest, or other effects that might have adversely influenced the results if a model had been used that did not take these into account. Different and perhaps unexpected results might be obtained by simulating (among other conditions) sets of items that poorly measure the ability levels of the population, by simulating items administered to distributions that are skewed, and by simulating situations where there are numerous missing data due to non-response rather than missing by design. Those wanting to investigate the effects or consequences of the different designs should conduct similar simulations with empirically obtained item parameters and ability distribution characteristics.

References

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model. A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 55–76). New York: Springer.

Barcikowski, R. (1972). A Monte Carlo study of item sampling (versus traditional sampling) for norm construction. *Journal of Educational Measurement*, 9(3), 209–214.

Beaton, A. E. (1987). *Implementing the new design: The NAEP 1983/1984 technical report*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*(2), 95–109.

Bock, R. D., Mislevy, R. J., & Woodson, C. (1982). The next stage in educational assessment. *Educational Researcher*, *11*(3), 4–11.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), pp. 39–53.

Gressard, R., & Loyd, B. (1991). A comparison of item sampling plans in the application of multiple matrix sampling. *Journal of Educational Measurement*, *28*(2), 119–130.

Johnson, E. J. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*(2), 95–110.

Johnson, M., & Lord, F. (1958). An empirical study of the stability of a group mean in relation to the distribution of test items among students. *Educational and Psychological Measurement*, *18*(2), 325–329.

Kleinke, D. (1972). *The accuracy of estimated total test statistics*. Washington, DC: National Center for Educational Research and Development. (ERIC Document Reproduction Service No. ED064356)

Knapp, T. (1968). An application of balanced incomplete block design to the estimation of test norms. *Educational and Psychological Measurement*, *28*, 265–272.

Lord, F. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*(2), 259–267.

Lord, F. (1965). *Item sampling in test theory and in research design* (ETS Research Bulletin No. RB-65-22). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational and Behavioral Statistics*, *8*(4), 271–288.

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & 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., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *(17)2*, 131–154.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *The NAEP 1983–84 technical report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mullis, I., Kennedy, A., Martin, M., & Sainsbury, M. (2006). *PIRLS assessment framework and specifications*. Chestnut Hill, MA: Boston College.

Mullis, I., Martin, M., Ruddock, G., O'Sullivan, C., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: Boston College.

Munger, G., & Loyd, B. (1988). The use of multiple matrix sampling for survey research. *Journal of Experimental Education*, *56*(4), 187–191.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data* [computer software]. Chicago, IL: Scientific Software.

Myerberg, N. J. (1975, April). *The effect of item stratification in multiple-matrix sampling*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Nair, K. R. (1943). Certain inequality relations among the combinatorial parameters of balanced incomplete block designs. *Sankhyā, 6*, 255–259.

National Center for Educational Statistics. (2010). *About NAEP*. Retrieved from http://nces.ed.gov/nationsreportcard/about/

Neidorf, T., & Garden, R. (2004). Developing the TIMSS 2003 mathematics and science assessments and scoring guides. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 23–66). Chestnut Hill, MA: Boston College.

Olson, J., Martin, M., & Mullis, I. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College.

Organisation for Economic Co-operation and Development (OECD). (2006). *Assessing scientific, mathematical, and reading literacy: A framework for PISA 2006*. Paris: Author.

Plumlee, L. (1964). Estimating means and standard deviations from partial data: An empirical check on Lord's sampling technique. *Educational and Psychological Measurement*, *14*(3), 623–630.

Preece, D. A. (1990). Fifty years of Youden squares: A review. *Bulletin of the Institute of Mathematics and Its Applications*, *26*, 65–75.

Pugh, R. (1971). Empirical evidence on the application of Lord's sampling technique to Likert items. *Journal of Experimental Education*, *39*(3), 54–56.

Reiser, M. (1983). An item response model for the estimation of demographic effects. *Journal of Educational Statistics*, *8*(3), 165–186.

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

Rubin, D. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 20−34). Alexandria, VA: American Statistical Association.

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

Scheetz, J., & Forsyth, R. (1977, April). *A comparison of simple random sampling versus stratification for allocating items to subtests in multiple matrix sampling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Shoemaker, D. M. (1970a). Allocation of items and examinees in estimating a norm distribution by item sampling. *Journal of Educational Measurement*, *7*(2), 123−128.

Shoemaker, D. M. (1970b). Item-examinee sampling procedures and associated standard errors in estimating test parameters. *Journal of Educational Measurement*, *7*(4), 255−262.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger Publishing Company.

van der Linden, W. & Carlson, J. (1999). *Calculating balanced incomplete block design for educational assessments*. Paper presented at the National Assessment Governing Board Achievement Levels Workshop, Boulder, CO.

van der Linden, W., Veldkamp, B., & Carlson, J. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement*, *28*, 317−331.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9−36.

von Davier, M. Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam, The Netherlands: Elsevier.

Yates, F. (1939). The recovery of inter-block information in variety trials arranged in three-dimensional lattices. *Annals of Eugenics*, *9*, 126–156.