ETS NAEP Technical and Research Report Series

Technical Report for the 2000 Market-Basket Study in Mathematics

John Mazzeo
Edward Kulick
Brenda Tay-Lim
Marianne Perie

February 2006
Technical Report
ETS-NAEP 06-T01

Listening.
Learning.
Leading.

THIS PAGE INTENTIONALLY LEFT BLANK.

ETS NAEP Technical and Research Report Series

Technical Report for the 2000 Market-Basket Study in Mathematics

John Mazzeo
Edward Kulick
Brenda Tay-Lim
Marianne Perie

February 2006
Technical Report
ETS-NAEP 06-T01

The work reported herein was done for the National Center for Education Statistics under the PR/Award No. R902F980001, CFDA No. 84.902F administered by the Office of Educational Research and Improvement, U.S. Department of Eduction. The authors would like to thank Robert J. Mislevy of the University of Maryland and Stephen L. Lazer of the Educational Testing Service, for numerous discussions about market-basket issues in general, for their work in helping to design the 2000 study, and for comments on this manuscript. The authors would also like to thank Wendy Yen and Dan Eignor of the Educational Testing Service, Susan Loomis of the National Assessment Governing Board, and Alex Sedlacek of the National Center for Education Statistics for the reviews of the manuscript. Lastly, we would like to acknowledge the work of Jeffrey Haberstroh of the Educational Testing Service who led the 1999 test development activities for the NAEP mathematics assessment.

THIS PAGE INTENTIONALLY LEFT BLANK.

Abstract

This technical report presents the goals and design of the 2000 National Assessment of Educational Progress (NAEP) market-basket study, describes the analyses that were conducted to produce the prototype NAEP market-basket report card, and presents and discusses results from the study that are pertinent to selected technical and psychometric issues associated with the potential implementation of a market-basket reporting option for NAEP. A *market basket* is a specific collection of test items intended to be representative or illustrative of a domain of material included in an assessment. Reporting assessment results in terms of the scores on this collection of items and publicly releasing the items are what is typically meant by *market-basket reporting*. Two market-basket test forms were constructed and administered to nationally representative samples of fourth-grade students. Results for a nationally representative sample of students from both sets of projections were compared with each other and with the results actually obtained by directly administering the market basket to separate nationally representative samples. While the two kinds of projection results were generally similar, differences between them, consistent with what one would expect from basic measurement theory, were evident. Furthermore, both sets of projection results were similar, in most cases, to actual results obtained by directly administering the market baskets to separate, randomly equivalent samples. There were, however, some notable differences.

Introduction

The National Assessment of Educational Progress (NAEP) is an ongoing assessment of what students in the nation and participating states and jurisdictions know and can do in a variety of academic subject areas (Horkay, 1999; see also http://nces.ed.gov/nationsreportcard/). The assessments are administered to representative samples of fourth-, eighth-, and twelfth-grade students, using a matrix-sampling design that assigns differing collections of items into separate test forms that are administered to different samples of students (see, for example, Allen, Carlson, & Donoghue, 2001). This design allows for the assessment of broadly defined content domains while minimizing the testing time—typically less than one hour—required of any individual examinee.

Due to the matrix sampling design used by NAEP, the data from the assessment are analyzed using Item Response Theory (IRT; see Lord, 1980; Mazzeo, Lazer, & Zieky, in press). Assessment results are reported primarily in terms of average scores and percents above cut points (PACs) for groups of students on IRT latent-variable scales. (Among the major NAEP reporting groups, for example, are students in the nation as a whole, public school students, private school students, male students, female students, White students, and Black students.)

The analysis of the NAEP mathematics assessment produced five distinct IRT-based scales (number properties and operations; measurement; geometry and spatial sense; data analysis, statistics and probability; and algebra and functions). The origin and unit size of IRT scales are arbitrary up to a linear transformation. For reporting purposes, the results for each of these content-area scales are linearly transformed from a within-grade metric that ranges from -4 to 4 to a cross-grade metric that ranges from 0 to 500 (see, for example, Jenkins, Chang, & Kulick, 1999). The primary reporting scale for NAEP mathematics results is a weighted average of the five transformed content-area scales, resulting in a cross-grade composite scale ranging from 0 to 500. Results for this composite scale are featured in NAEP report cards (see, for example, Braswell, Lutkus, Grigg, & Santapau, 2001). For major reporting groups, composite score averages and PACs are the “coin of the realm” of NAEP reporting.

Over the years, the NAEP program has tried various ways to make the behavioral meaning of its latent-variable scale units accessible to the general public. The use of scale

anchoring and anchor-level descriptions in long-term trend NAEP (Beaton & Allen, 1992), achievement-level reporting and achievement-level descriptions in main NAEP (National Assessment Governing Board [NAGB], 1995), exemplar items and item maps (Braswell et al., 2001, pp. 115–138) all represent attempts to make the NAEP results more meaningful and easier to interpret for the public.

In 1996, NAGB redesigned NAEP (NAGB, 1996) and generated considerable interest in using *market-basket reporting* (DeVito & Koenig, 2000; Forsyth, Hambleton, Linn, Mislevy, & Yen, 1996; Mislevy, 1998, 2000) to make NAEP results meaningful to the public. A *market basket* is a specific collection of assessment items intended to be representative or illustrative of a domain of material included in an assessment. Market-basket reporting, as it is discussed in the references listed above, refers to reporting assessment results in terms of scores on this specific collection of items, along with the release of those items to the public.

As part of the 2000 NAEP mathematics assessment, a study was conducted to pilot test a market-basket reporting option for NAEP. The study, conducted at grade 4 in mathematics, consisted of constructing two market-basket test forms, along with a number of additional NAEP forms required to achieve the analytic goals of the study, and administering them to nationally representative samples of students. The data from this study served two purposes: to provide a basis for producing a prototype NAEP report card based on market-basket reporting, and to provide data for studying selected technical issues associated with the potential implementation of a market-basket approach in NAEP.

This report first describes the design of the 2000 market-basket study and then discusses the analyses that were carried out to produce the prototype NAEP market-basket report card. The next section of this report describes efforts at designing a short-form market-basket approach. Such an approach would allow states or districts to administer a particular market-basket form to a group of students in years when NAEP is not administered, as a lower-burden alternative or as a special lower-cost option for districts that wanted to participate in NAEP. The final section of this report explores certain technical issues surrounding the implementation of a market-basket reporting and design option in NAEP.

Before proceeding, it is important to distinguish between two potential uses of a market-basket approach in NAEP: market-basket *reporting* of the results from a full NAEP administration, which includes many blocks of items distributed across multiple forms; and the use of market-basket *test forms* to collect assessment data as well as to report assessment results.

Market-Basket Reporting

As noted above, the term *market basket* refers to a specific collection of assessment items. The term *market-basket reporting* typically refers to the expression of assessment results as simple percent-correct scores on a collection of items, along with the public release of those items. The use of market-basket reporting, with some variations, has been proposed by a number of researchers.

In a paper on domain-referenced scoring, Bock (1996) suggested using a large collection of items (anywhere from 500 to 5,000), effectively providing an operational definition of the skill domain in question. Bock advocated public release of the entire domain of items to stimulate discussion and learning in the subject area. A second proposal, by Mislevy (2000), involved the use of what he refers to as a *synthetic* market-basket form. Such a market basket would be large enough to convey adequately the mix of formats, skills, and topics specified by the framework for the assessment, but would not be exhaustive. The synthetic market basket, as a whole, could be larger than the test forms administered to any individual test taker, but would not consist of the entire item pool. A third proposal (Mislevy, 1998; 2000) entailed constructing market baskets equivalent in length to the actual test forms administered to examinees as part of the assessment. Mislevy (2000) refers to these as market-basket *administrable forms*.

A key point about market-basket reporting is that it is possible to report results for students regardless of whether the students have been administered all or part of the market basket. Thus, in NAEP, it is possible to consider market-basket reporting within the context of the current design of matrix-sample data collection, by identifying a subset of items that, as a whole, is proportionally equivalent to a single NAEP test form and that could be released as part of a market-basket approach. This design would not require that these items actually be assembled as a single test form at the time of administration, or

that they be administered as an intact instrument to the sample from which the market-basket results would be derived. As discussed further below, some supplemental data collection, along with additional complex IRT-based analyses, may be required for the best implementation of market-basket reporting in the current matrix-sample design. However, converting to market-basket reporting in NAEP would require only the re-expression of the current 0 to 500 NAEP scale scores and achievement-level results in terms of percent-correct market-basket scores. The section on producing a prototype market-basket report, below, describes the analyses required to implement such an approach and provides an illustration of market-basket results, using the data from the NAEP assessment in mathematics at grade 4.

Short-Form Market-Basket Data Collection

A second potential configuration for the program, one that generated considerable discussion at the time of the 1996 NAGB redesign of NAEP, would combine the use of market-basket reporting with a short-form data collection option. With this option, main NAEP would proceed as it is currently designed (i.e., data are collected with a matrix-sampling design and results are reported in terms of marginal estimates of reporting-group statistics on latent-variable scales), while also offering market-basket short forms as alternatives. These short forms could be administered to some subset of the state and national samples during standard NAEP reporting cycles, to provide the basis for preliminary fast-track reporting of results (i.e., reporting based on simpler analysis methods than those associated with NAEP's matrix-sample design). Alternatively, they could be offered to states for use in nonassessment years as a lower-burden alternative or as a special lower-cost option for districts that wanted to participate in NAEP. Sections 2 and 3 below discuss and illustrate some of the potential complexities involved in implementing such an option, using data from the 2000 NAEP market-basket study.

A Description of the 2000 NAEP Market-Basket Study

The study was designed to address three specific goals:

- to gain experience in defining and constructing market-basket forms for reporting or data collection purposes,

- to produce and evaluate a prototype market-basket report of NAEP results based on current matrix-sample data collection procedures,
- to conduct research on selected methodological and technical issues that might arise if a short-form market-basket data collection were to be implemented.

Gaining Experience in Defining and Constructing Market-Basket Forms

Before the 2000 assessment, the development of market-basket forms in NAEP was discussed at an abstract level. Content frameworks and test specifications for developing assessment item pools were in place for the NAEP subject areas, but there were no analogous documents to guide the definition and construction of market baskets. Thus, one goal of the 2000 study was to gain concrete experience in defining and constructing market baskets in NAEP, providing guidance to NAGB and NCES for developing policies and procedures to guide future market-basket work in mathematics and other subject areas.

Producing and Evaluating a Prototype Report of NAEP Results

The hypothesized utility of market-basket reporting lies in the presumption that results presented in this fashion would be more meaningful for, and more accessible to, the public interested in education. Currently, NAEP reports results on latent-variable scales derived through the use of IRT. Such results look different from the percent-correct scores that the public is used to seeing on classroom tests. A market-basket reporting option could make NAEP scores more accessible for the public by defining the NAEP scale in a way that is presumably more familiar, i.e., simple percent-correct[2] scores on a particular collection of items (in this case, the market basket). Moreover, this set of items could be generally available for public examination, much like the tests students bring home from school to show their parents.

[2] The term "percent-correct" is most appropriately used for tests that consist entirely of items that are scored right or wrong. A NAEP Market Basket is likely to consist of a mix of items scored as either right/wrong or scored by providing from partial to full credit. Thus, strictly speaking, total scores are more properly described as "percent of total available points." For ease of exposition, the term "percent-correct" is used throughout this paper as shorthand for "percent of total available points."

Despite the obvious intuitive appeal of market-basket reporting, before this study NAEP had not produced an example of what such a report might look like. Without a concrete example, it was difficult to evaluate the degree to which market-basket reporting would actually make NAEP results more meaningful. Thus, a second key goal of the 2000 study was to produce and evaluate a prototype market-basket report card in NAEP.

Conducting Research on Selected Methodological and Technical Issues

As noted earlier, the implementation of a market-basket reporting system in NAEP could be combined with short-form market-basket data collection. In such a configuration, results would typically be obtained by administering the full NAEP assessment instrument with its usual use of matrix sampling and complex IRT analyses. The typical NAEP scale score results derived under this design would be re-expressed as market-basket scores. However, results for some jurisdictions or in some assessment cycles might be obtained by directly administering a market basket form to each student in the relevant samples and directly calculating each student's percent-correct scores. These student scores might then be aggregated and the aggregated scores compared to the "re-expressed" NAEP results obtained under the typical NAEP design. The implementation of such a hybrid system entails policy, development, and analytic issues that needed to be addressed. (See, e.g., Mislevy [2000] for a discussion of many of these issues.)

A key issue inherent in the Mislevy discussion is whether and how the two sets of results—NAEP market-basket results projected from the matrix-sample design and results collected directly with the market basket—can be made sufficiently comparable to support the program's goal of accurate and valid comparisons of state and national results, within each assessment year and over time. As just described, the use of a market-basket *data collection* in NAEP, if it achieved the goal of simplifying matters, would involve the production of individual student-level market-basket scores and the aggregation of these scores to obtain group-level results. In contrast, current NAEP estimates of average scores and PACs (i.e., estimates obtained through the matrix-sample design, IRT analysis, and marginal estimation) are derived from estimates of the distribution of student proficiencies expressed on latent-variable scales.

At least two options exist for converting the latent-variable estimates to a market basket scale. The simplest option (referred to by Mislevy [2000] as *one-stage projection*)

converts the latent-variable results to what is commonly referred to in the educational measurement literature as a market-basket percent-correct *true-score* scale (see, e.g., Lord & Novick, 1968). True scores can be thought of, conceptually, as the scores that students would obtain if it were possible to administer an extremely long test—in the current context, a very long test consisting of items like those in the market basket. If NAEP were considering only a market-basket reporting option, one-stage projection (which is described and illustrated below with the data from the current study) would be the obvious choice for the conversion.

Administering the market-basket short form produces what are commonly characterized in the measurement-theory literature as *observed scores* (Lord & Novick, 1968.) Because an actual market-basket test consists of a sampling of the items that would make up the hypothetical *long* market basket test, the observed scores and hypothetical true scores of individual students are not the same. The difference between these two types of scores at the level of the individual is what is typically referred to by the measurement community as *measurement error*. A key result of measurement theory that is of direct relevance to the current discussion is that the distribution of observed scores for a group of examinees is generally not the same as the distribution of their true scores, due primarily to the impact of measurement error on the former. Hence, NAEP estimates of group statistics based on aggregation of observed scores from a market-basket short form will not necessarily align themselves well with NAEP estimates of true-score quantities. The two sets of estimates are, in a sense, on different scales, and are estimating somewhat different quantities. The degree to which the two sets of estimates differ is an empirical matter. Generally speaking, the differences will in large part be a function of the length of the market-basket short form. The data from the current study is used below to examine to some degree the extent of the differences between the two types of estimates and to explore ways to carry out the re-expression of NAEP scale score results as market-basket scores.

The remainder of this report is divided into three sections. The first section describes the design of the market-basket forms, the student sample, and the booklet spiral, and presents some basic results obtained from the administration of the forms. The second section discusses the steps involved in producing the prototype market-basket

report and briefly discusses the main NAEP results re-expressed as market-basket scores. The third section compares the results obtained by directly administering the market-basket short form to those obtained by projecting the main NAEP results into market-basket reporting metrics. Two types of projections are considered: the one-stage projection used for the market-basket report, and a two-stage projection that converts latent-variable results to market-basket observed-score results.

Section 1: Study Design

Defining the Market Basket

A market basket of NAEP mathematics items was defined as a collection of items representative of the full domain covered by the grade 4 mathematics assessment. The full NAEP item pool for any given assessment year is designed to meet the content, process, and item format specifications laid out in NAGB framework documents. Therefore, the 2000 study used the existing mathematics item pool to provide a set of specifications for defining the market basket. Specifically, the 1996 mathematics item pool for grade 4 was the target in this study, and a market basket was defined as a collection of items that matched the total item pool in content, process, format, and statistical characteristics.

The market-basket test forms were designed to be comparable in length to current NAEP test forms (i.e., they could be administered in a 45-minute testing window) for three reasons. First, while states and districts had expressed an interest using short forms as a low-cost data collection surrogate for the full NAEP (perhaps in off years), they consistently expressed concerns over testing burden. It seemed unlikely that they would accept a short form that was any longer than the current NAEP assessment forms.

Secondly, market-basket reporting presumes the release of a market basket for public review. Meaningful public review is less likely to occur if the market basket consists of large numbers of items (see Bock, 1996, for an alternative point of view).

Thirdly, in order to maintain the size and composition of future mathematics item pools, NAEP has to develop extra items to be released as market-basket items or to replace them. In order to control test development costs, it seemed reasonable to limit the number of items in the market basket to that in a typical NAEP test form.

Constructing Market-Basket Forms

Two forms were created: one to be kept secure for future market-basket reporting options (MB1) and the other to be released in the prototype report (MB2). Although every effort was made to make these forms as parallel as possible, the nature of the items and the specifications they assessed resulted in two forms that covered the same content but that were slightly different in length.

Exhibit 1 displays characteristics—the specifications and statistical properties—of the item pool for the 1996 NAEP mathematics assessment and for the two market-basket forms. The 1996 NAEP mathematics item pool consisted of items from five *content* areas (number properties and operation; measurement; geometry and spatial sense; data analysis, statistics and probability; and algebra and functions) and three *process* areas (procedural knowledge, conceptual understanding, and problem solving). Items were of three *formats*: multiple-choice items, constructed-response items that are scored correct or incorrect, and constructed-response items that are scored on a multipoint scale with potential scores ranging from 0 to *k*, where *k* is the maximum score of the item. Of the multipoint items, some require a brief response from the examinee and are scored on a 3-point scale (no credit, partial credit, full credit), some require a more extended constructed response and are scored on a 5-point scale (no credit, minimal, partial, satisfactory, complete), and some are sets of multiple choice items that are scored collectively.

The NAEP mathematics test development committee was given the task of identifying a set of secure NAEP items for the market basket from the existing 1996 NAEP item pool. They were to be high quality exemplars of the pool; to match the NAEP item pool in content, process, format, and statistical specifications; and to be possible to administer collectively as a 45-minute NAEP form.

In exhibit 1, column 1 presents the percentages of items specified for the 1996 grade 4 mathematics assessment by content, process, and item format. These percentages served as content specifications for the market-basket forms. The bottom of column 1 presents the statistical characteristics of the 1996 item pool, specifically, the mean and standard deviation of Classical Test Theory (CTT) item difficulty indices, expressed on the ETS *delta scale* and the *mean r-biserial* or *item discrimination index*. These statistics served as statistical specifications for the market-basket forms. (For information about these statistics, see Hecht & Swineford, 1981.)

In exhibit 1, column 2 shows the makeup of the 33-item market-basket collection selected by the committee, in terms of number of items for each content, process, format, and statistical specification. For comparison purposes, column 3 shows the target test specifications expressed in terms of the expected number of items in a 33-item form. The

goal for developing the market-basket forms was to match the target percentages shown in column 1. The actual percentages were very close: the 33-item test content did not differ from the 1996 target percentages by more than one percentage point for any of the five content areas. The statistical characteristics of this newly constructed form, shown in the bottom portion of the exhibit, are based on data for these items from the 1996 assessment (column 2), which may be compared with the statistical characteristics of the entire 1996 item pool (column 3). Following standard NAEP practice at that time, these 33 items were organized into three distinct blocks, each of which was designed to be administered in a separately timed 15-minute period.[3] For the remainder of this report, this market-basket test form is denoted as MB1, and its 3 blocks are denoted as M26, M27, and M28.

[3] Beginning with the 2002 NAEP assessment, test booklets for all NAEP subject areas, including mathematics, consist of two 25-minute blocks of subject-matter items.

Exhibit 1. Market-basket test specifications

		Market-basket forms			
		MB1: 33-item form		MB2: 31-item form	
Content specification:	Target: 1996 percentage	Actual number of items	Target number of items	Actual number of items	Target number of items
Number sense	40%	13	13	12	12
Measurement	20%	7	7	6	6
Geometry	15%	5	5	5	5
Data analysis	10%	3	3	3	3
Algebra	15%	5	5	5	5
Total	100%	33	33	31	31
Process specification:					
Procedural knowledge	33%	10	11	10	10
Conceptual understanding	33%	11	11	10	10
Problem solving	33%	12	11	11	10
Total	100%	33	33	31	31*
Item format:					
Multiple-choice	55%	19	18	17	17
Binary constructed-response	24%	7	8	7	7
Partial credit constructed-response	21%	7	7	7	7
Total	100%	33	33	31	31
Statistical specifications:					
Mean item difficulty (Delta)	13.2	12.8	13.2	13.2	13.2
SD of item difficulties	2.3	1.9	2.3	2.2	2.3
Mean *r*-biserial	0.63	0.62	0.63	0.62	0.63

*The target was equal number of items per process specification. Because there were three categories resulting in 31 items, one category would need an extra item, but no particular category was specified a priori.

Using form MB1 as a model, the mathematics test-development committee produced six new blocks of items. The items in these blocks were designed to measure content and outcomes similar to those in MB1. These six blocks of items were field tested in 1999, and, based on field test results, the committee assembled from this pool what they judged to be the best three-block, 45-minute form possible that met the market-basket specifications. This second form (subsequently denoted as MB2) was assembled to the same test specifications, but consisted of three blocks of newly developed items (subsequently denoted as M23, M24, and M25). MB2 was intended for release as part of the 2000 prototype market-basket form, while MB1 will be retained as a secure market-basket form for use in future assessments. Exhibit 1, column 4 shows the makeup of this 31-item market-basket collection, and column 5 shows the target test specifications for a 31-item form. The statistical characteristics of this newly constructed form are based on data for these items from the 1999 pilot test (column 4), which may be compared with the statistical characteristics of the entire 1996 item pool (column 5).

Although MB2 appears to meet the market-basket specifications exactly, it differs slightly in length from MB1. Given the six blocks of newly field-tested items available, the mathematics test development committee was not able to arrive at a 33-item test that it felt met all the necessary specifications, contained only exemplary items, and could be realistically administered in the 45 minutes. They chose, instead, to produce a slightly shorter market-basket form. Further, the item statistics for this shorter form are based on 1999 field-test samples and are not necessarily comparable to those shown for MB1, which are based on 1996 NAEP operational samples.

Student Samples and Booklet Spirals

The 2000 NAEP mathematics assessment at grade 4 was administered to two distinct, though randomly equivalent, school and student samples, each of which was assigned a different collection of test booklets. One, the main NAEP sample, on which the official NAEP mathematics results are based, consisted of 13,511 students. This sample was administered the main NAEP assessment instrument, which consisted of 13 blocks of items arranged according to a balanced incomplete block (BIB) design into 26 distinct test booklets. Each test booklet contained three separately timed 15-minute blocks. The

test booklet design for main NAEP is shown in exhibit 2. Although individual items from MB1 appear in various places throughout the 13 main NAEP blocks, these items do not appear in the main assessment as the intact blocks M26, M27, and M28.

The other sample in the 2000 assessment was a "market-basket sample." This sample of 8,012 students, drawn from a different set of schools from the ones that participated in the main NAEP assessment, allowed for the IRT calibration of the market-basket forms and provided the data necessary to investigate a number of methodological and analytic issues. The test booklet design for the market-basket sample is given in exhibit 3.

Each student in the market-basket sample was administered one of seven test forms. Forms MB1 and MB2 are the market-basket forms described above. The remaining forms were used to link MB1 and MB2 with one another and to the main NAEP scale. Even though all of the items in MB1 were taken from the main NAEP assessment, they were not configured into equivalent blocks in the main assessment. Therefore, items selected for MB1 and administered in equivalent blocks were treated the same as the new items in MB2; both sets of items needed to be linked to the main NAEP scale. The forms denoted LINK1 and LINK2 were hybrid forms used to calibrate the two market-basket forms to a common scale. Each consists of two blocks of items from one of the market-basket forms and a third block from the other market-basket form. The final three forms (LINK3, LINK4, and LINK5) were used to calibrate the new market-basket forms to the main NAEP scale. Each form consisted of one block from MB2 paired with two intact blocks from the main NAEP BIB. The purpose of these linking forms was to permit the calibration of the seven market-basket forms and the 26 main NAEP test forms to a common scale.

Exhibit 2. Booklet design for main NAEP

Booklet	Block 1	Block 2	Block 3	Sample size
Main1	M3	M4	M7	529
Main2	M4	M5	M8	513
Main3	M5	M6	M9	522
Main4	M6	M7	M10	525
Main5	M7	M8	M11	513
Main6	M8	M9	M12	508
Main7	M9	M10	M13	502
Main8	M10	M11	M14	510
Main9	M11	M12	M15	505
Main10	M12	M13	M3	506
Main11	M13	M14	M4	512
Main12	M14	M15	M5	516
Main13	M15	M3	M6	520
Main14	M3	M5	M10	533
Main15	M4	M6	M11	508
Main16	M5	M7	M12	523
Main17	M6	M8	M13	531
Main18	M7	M9	M14	526
Main19	M8	M10	M15	520
Main20	M9	M11	M3	520
Main21	M10	M12	M4	537
Main22	M11	M13	M5	540
Main23	M12	M14	M6	522
Main24	M13	M15	M7	521
Main25	M14	M3	M8	529
Main26	M15	M4	M9	520
			Total	13,511

Shading denotes a linking block between main NAEP and the market-basket spiral.

Note: Thirty-three of the items in main NAEP blocks M3 to M15 also appear in the market-basket spiral as blocks M26, M27, and M28.

Exhibit 3. Booklet design for the market-basket sample

Booklet	Block 1	Block 2	Block 3	Sample size
MB1	M26	M27	M28	1,976
MB2	M23	M24	M25	2,033
LINK1	M23	M27	M25	991
LINK2	M26	M24	M28	1,012
LINk3	M11	M24	M14	687
LINK4	M13	M9	M25	661
LINK5	M23	M4	M8	8,012

Shading denotes a linking block between main NAEP and the market-basket spiral.

Note: Thirty-three of the items in main NAEP blocks M3 to M15 also appear in the market-basket spiral as blocks M26, M27, and M28.

Results

Table 1 presents some simple percent-correct statistics for the market-basket forms. The goal in constructing the two market-basket forms was to produce two parallel versions of the market-basket test. The results in Table 1 suggest that, some, though not full, success in achieving that goal. MB1 appears to be a slightly easier form than MB2, in that the average percent-correct score for the samples administered MB1 was about 1.5 points higher than the sample administered MB2. This overall difference, however, is not statistically significant. The standard deviations of scores on the two forms were about the same, as were the coefficient alpha estimates (Lord & Novick, 1969) of the reliability coefficients, suggesting quite similar degrees of measurement precision for the market-basket forms, despite the modest difference between them in test length.

Table 1. Actual observed-score percent-correct statistics for market-basket study samples on forms MB1 and MB2

	MB1					MB2				
	Number of students	Percent of students	Mean	Standard deviation	Reliability	Number of students	Percent of students	Mean	Standard deviation	Reliability
Total	1,976	100.0 (0.0)	47.1 (0.5)	19.5	0.87	2,033	100.0 (0.0)	45.6 (0.6)	19.6	0.86
Gender										
Male	973	50.6 (1.3)	46.6 (0.8)	19.6		1,001	50.4 (1.3)	46.5 (0.8)	19.6	
Female	1,003	49.4 (1.3)	47.6* (0.8)	19.5		1,032	49.6 (1.3)	44.6 (0.7)	19.6	
Male - Female			-1.0					1.9		
Race/Ethnicity										
White	1,103	66.0 (0.4)	51.8 (0.6)	18.6		1,108	66.0 (0.4)	50.1 (0.8)	19.1	
Black	359	14.3 (0.3)	34.2 (0.9)	16.0		366	14.0 (0.3)	32.4 (1.2)	15.2	
Hispanic	408	15.0 (0.3)	37.3 (1.0)	16.6		419	15.1 (0.3)	36.9 (0.9)	16.5	
White - Black			17.6					17.7		
White - Hispanic			14.5					13.2		
Type of school										
Public	1,746	88.9 (1.1)	46.1 (0.6)	19.4		1,800	88.6 (1.1)	44.8 (0.6)	19.4	
Nonpublic	230	11.1 (1.1)	54.5 (1.7)	18.9		233	11.4 (1.1)	51.2 (2.4)	20.0	
School lunch eligibility										
Eligible	775	34.1 (1.5)	35.5 (0.8)	16.1		789	33.2 (1.4)	35.8 (0.7)	16.5	
Not eligible	1,201	65.9 (1.5)	53.0* (0.7)	18.4		1,244	66.8 (1.4)	50.4 (0.7)	19.2	

*Statistically significantly different from MB2.

NOTE: The standard errors of the statistics in the table appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Table 1 also presents the results disaggregated by major NAEP reporting groups. In general, scores on MB1 appeared higher than those on MB2 for most subgroups, with two notable exceptions—for groups defined by gender and for groups defined by eligibility for free or reduced-price school lunch. Average scores for males on the two forms were nearly identical, while average scores for females were almost 3 points higher on form MB1 than on form MB2. As a result, the average score difference between males and females went in opposite directions on the two forms. On MB2, males scored higher than females, as has traditionally been the case in NAEP mathematics assessments (see, e.g., Braswell et al., 2001). The reverse was true on form MB1, though this difference between males and females was not statistically significant. Similarly, the average score of those eligible for free or reduced-price school lunch was nearly identical on the two forms, while the average score for those who were not eligible was higher on MB1 than on MB2.

Section 2: Producing the Prototype Market-Basket Report

The results from MB2 were used to construct a prototype market-basket report by converting the national scale-score results for grade 4 to a market-basket true-score metric. This required two analysis steps: IRT calibration of the items in the market-basket short forms to the main NAEP reporting scale and conversion of NAEP scale scores to the market-basket true-score metric. The results were then presented as average percent correct overall and for subgroups, and as percentages of students above cut scores expressed as percent correct on the market-basket form.

IRT Calibration of Market-Basket Test Forms

Conversion of main NAEP results to either of the market-basket forms required a set of IRT item parameters, calibrated to the existing NAEP mathematics scales, for each of the items contained in those forms. As discussed in more detail below, the existence of such a set of item parameters provides a scoring function, denoted here as $v(\bullet)$, as well as estimates of $P(\underline{x} \mid \underline{\theta})$, both of which can be used to convert NAEP scale-score results to a market-basket reporting metric.

To obtain estimates of IRT item parameters for the market-basket forms, the data from the main NAEP national assessment (i.e., the 26-book matrix-sample instrument) and the market-basket study (i.e., the seven-book spiral containing the two market-basket forms and five linking forms) were pooled, and a *concurrent calibration* of all 33 test forms (i.e., the 26 test forms in the main NAEP assessment and the seven test forms in the market-basket study) was conducted. For this special calibration, item parameters for all intact blocks from the main NAEP instrument were held fixed at the values estimated during the NAEP operational calibration of the grade 4 year-2000 assessments (see the technical report on the 2000 mathematics assessment at the assessment procedures section of the NAEP website at http://nces.ed.gov/nationsreportcard/tdw).[4] Thus, the special calibration produced item parameter estimates only for blocks M23 through M28.

[4] There are alternative approaches to calibrating the market-basket results to the main NAEP scales. We chose a "fixed-parameter" approach for a number of practical and theoretical reasons. A more detailed discussion of the possible approaches and a comparison of results based on these alternatives can be found in Tay-Lim, Wang, & Kulick (2002).

The presence in the market-basket study of intact *linking* blocks common to the main assessment, and the fixed item-parameter estimates for these blocks at their operational values made it possible to calibrate the items in MB1 and MB2 to the existing operational scales. Furthermore, the two hybrid forms (LINK1 and LINK2) consist of blocks of items from both market-basket forms, which allowed the two sets of market-basket items to be calibrated to those same operational scales.

The scaling of the market-basket forms was carried out using the NAEP BILOG/PARSCALE program.[5] The program uses an E-M algorithm to obtain *marginal-empirical-Bayes-model* and *marginal-maximum-likelihood* estimates of item parameters and the parameters of the marginal proficiency distribution of θ, respectively. Two models for the distribution are available: a *normal distribution* model, and a *nonparametric* model, in which θ is modeled as a discrete variable with a multinomial distribution over a fixed set of points.[6] For the special market-basket calibration, the marginal proficiency distribution for the main NAEP sample was held fixed at the values estimated from the operational calibration, while a nonparametric estimate of the proficiency distribution for the market-basket sample was obtained concurrently with the estimates of the market-basket item parameters. It is important to note that the BILOG/PARSCALE proficiency distribution estimates (main NAEP and market-basket sample) played no further role in the market-basket analysis. As discussed below, only the item-parameter estimates from these calibrations were used in the subsequent analyses reported here.

Five separate content-area subscales (number properties and operation; measurement; geometry and spatial sense; data analysis, statistics and probability; and algebra and functions) were produced operationally, and the items from market-basket blocks were calibrated to these existing scales. As is done for the main NAEP assessments, each item was constrained to load on one of the five content-area scales. Within each content area, items from different process categories and different formats were combined on a single scale using the PARSCALE program.

[5] This is a special ETS-produced, combined version of the BILOG and PARSCALE programs, which shares many features with commercially available versions of BILOG and PARSCALE, but also has a number of different capabilities, as well as an extended set of diagnostics. ETS has a license from Scientific Software to produce this combined version solely for its own use.

Converting NAEP Scale Scores to the Market-Basket Metric

One of the main goals of the 2000 study was to provide an example of how NAEP results (i.e., average scores and percentages at or above achievement levels for the sample overall and for major reporting groups such as those defined by gender and race/ethnicity) would look were they to be expressed in a market-basket metric. In order to accomplish this goal, the previously published 2000 NAEP national scale-score results for grade 4 (Braswell et al., 2001) were converted to NAEP market-basket results. Two types of conversions were performed—conversion to a market-basket true-score metric (which is discussed in this section) and conversion to a market-basket observed-score metric (which is discussed in section 3 of this report). Only the true-score results were shown in the prototype report, and discussion in this section focuses on the methods and results involved in producing these.

The operational analysis on which the 2000 NAEP scale-score results were based (i.e., the results reported in Braswell et al., 2001) resulted in a set of five *plausible values* (PVs) on each of the five mathematics content-area subscales for each examinee in the national sample. These PVs are random draws from each individual's conditional posterior distribution of scale scores, given responses to the items in the test, standing on a large number of demographic and instructional variables, a set of estimated IRT item parameters, and estimated parameters for a model relating the demographic and background characteristics to NAEP scale scores. The PVs are intermediate values used to calculate estimates of scale score distributions, i.e., estimates of $f(\underline{\theta})$, scale score means, standard deviations, PACs, and associated standard errors. These estimates are calculated for the overall NAEP samples and for a large number of demographic groups and other reporting variables. A detailed explication of NAEP analysis procedures is beyond the scope of this paper, but see Mislevy (1991) as a basic reference; Allen, Carlson, Johnson, & Mislevy (2001) for a more current explication in the context of NAEP; and the technical report on 2000 Mathematics (found in the assessment procedures section of the NAEP website).

---

[6] In NAEP, the multinomial model is routinely used with the points being 41 equally spaced θ -values between -4 and +4.

The PVs from the main NAEP operational analysis, along with the estimates of IRT item parameters for the market-basket forms, were used to obtain estimates of $G(\tau)$ and $H(x)$ (i.e., the cumulative density functions of the market-basket true scores and observed scores, respectively) and the corresponding NAEP means, standard deviations, PACs, and standard errors implied by these estimated distributions. The key step in this process was the conversion of the existing PVs into *plausible market-basket scores*. Once this conversion was accomplished, the standard NAEP analysis procedures, described in Allen, Carlson, Johnson, et al. (2001), were applied to the transformed scores to obtain estimates in the true-score metrics of MB1 or MB2.

Obtaining Market-Basket True Scores

As is well known (see, for example, Lord, 1980), IRT item-level true scores are typically defined as the expected item score given θ (i.e., $E(X_{jk} \mid \theta)$). Expected scores on the market-basket tests (i.e., market-basket true scores) were derived directly from the PVs associated with the operational NAEP sample and the estimated item parameters from the special market-basket IRT calibration. Expected item scores were summed within each subscale, and these subscale sums were then added across all subscales and subsequently converted to a percent-correct metric. Specifically, if τ_{ip} denotes the market-basket true score (in "percent-correct terms") for individual *i* based on plausible value *p*, then

$$\tau_{ip} = \nu(\underline{\theta}_{ip}) = \frac{100}{M}\sum_{k=1}^{5}\sum_{j:k} E(X_{jk} \mid \theta_{ikp}) = \frac{100}{M}\sum_{k=1}^{5}\sum_{j:k}\sum_{l=0}^{M_{jk}} lP_{jkl}(\theta_{ikp}),$$

where

θ_{ikp} = plausible value *p* of student *i* on subscale *k*,

$P_{jkl}(\theta)$ = the estimated probability of observing a score l on item *j* from subscale *k*, conditioned on θ, which came from the special market-basket IRT calibration described above,

$M_{jk} =$ the maximum score on item *j* from scale *k*, and

$M = \sum_{k=1}^{5}\sum_{j:k} M_{jk}$, the maximum possible score over all items.

Like the PVs from which they were converted, the sets of true scores (τ_{ip}) are intermediate values that were used to calculate estimates of market-basket true-score means, PACs, and distributions, as well as their associated standard errors, for the overall NAEP samples and for the large number of demographic groups and other kinds of student groups on which NAEP typically reports.

Determining Achievement-Level Cut Points in the Market-Basket Metric

NAEP results are also reported in terms of achievement levels—specifically, the percentages of students performing at or above each of three NAEP cut scores on a NAEP scale. (For mathematics assessments, the scale is a composite of the subscales previously mentioned.) The regions on the NAEP composite-score scale implicitly defined by the three cut scores are labeled *Basic, Proficient,* and *Advanced,* with the area falling below the cut score for *Basic* labeled "below *Basic*."

In order to examine achievement-level results as expressed in a market-basket metric, the achievement-level cut scores, which are typically expressed in the NAEP composite scale-score metric, were converted to the metric of market-basket scores. An *equipercentile* equating method was used to define market-basket achievement-level cut scores. In the equipercentile method, the percentage of students in the 2000 national reporting sample who performed above the cut scores for the three achievement levels—*Advanced, Proficient,* and *Basic*—was calculated. Then the projected market-basket true scores for the same 2000 national reporting sample were rank-ordered from low to high, and the percentiles were located that corresponded to the percentage of students at each achievement level in main NAEP. These percentiles were used as the achievement-level cut scores in the market-basket true score metric.

Results for the Prototype Market-Basket Report

In the prototype report, the average number of points scored on market-basket form MB2 was reported for all fourth-graders in the nation, and then for student groups defined by gender, race/ethnicity, type of school, eligibility for free or reduced-price school lunch, and Title I participation. In addition, the percentage of students performing at each of the three achievement levels—*Basic, Proficient,* and *Advanced*—as well as the percentage

performing below *Basic* was reported for each student group. Then, each of the 31 items in the MB2 form was displayed, ordered in terms of student performance. That is, the item that students, as a whole, performed best on was shown first, and the item that the students performed least well on was shown last. For each multiple-choice item, the full question and response options were given, along with the correct response and the percentage of students answering the item correctly. In addition, the percentage correct for students within each achievement level was provided. For the constructed-response items in the market basket that were scored as correct or incorrect, the prompt and any graphics were provided, and the percentage of all students answering the item correctly was given along with the percentage correct for students within each achievement level. For the constructed-response items that were scored with partial credit, the percentage of all students providing either "complete" responses (for items scored with three levels) or "substantial" or higher responses (for the item with five levels) was given. Samples of these responses were shown. Also, the corresponding percentages for students within each achievement level were provided.

Descriptive Statistics Shown in the Prototype Report

Table 2 presents NAEP grade 4 composite scale-score results, as reported in Braswell, et al. (2001), along with the projected MB2 true-score results produced for the prototype market-basket report using the methods described above. All results were based on the sample of students that took the main NAEP matrix-sample assessment. The average fourth-grader scored 228 on the NAEP composite mathematics scale, which ranged from 0 to 500. This scale score corresponded to a percent-correct score of 47 percent on MB2. Patterns of subgroup differences in terms of market-basket scores reflected the same patterns of subgroup differences evident in the "official" NAEP composite scale-score results.

Table 2. Projected true-score percent-correct statistics for the main NAEP sample on form MB2 and actual Main NAEP composite scale scores

			Main NAEP		MB2	
	Number of students	Percent of students	Mean	Standard deviation	Mean	Standard deviation
Total	13,511	100.0 (0.0)	227.6 (0.9)	31.2	46.8 (0.5)	19.1
Gender						
Male	6,680	50.7 (0.7)	228.9 (1.0)	32.2	47.8 (0.6)	19.6
Female	6,831	49.3 (0.7)	226.3 (0.9)	30.0	45.9 (0.5)	18.5
Race/Ethnicity						
White	8,581	65.9 (0.3)	235.7 (1.0)	27.9	51.8 (0.6)	18.2
Black	1,795	14.2 (0.2)	205.1 (1.6)	27.6	33.1 (0.7)	14.0
Hispanic	2,239	15.0 (0.3)	211.5 (1.5)	30.6	37.2 (0.8)	16.4
Type of school						
Public	7,070	89.0 (0.5)	226.2 (1.0)	31.4	46.0 (0.6)	19.1
Nonpublic	6,441	11.0 (0.5)	238.5 (0.8)	26.6	53.4 (0.5)	17.8
School lunch eligibility						
Eligible	3,353	32.2 (1.0)	209.7 (1.0)	29.5	36.0 (0.5)	15.6
Not eligible	10,158	67.8 (1.0)	236.1 (1.0)	28.2	52.0 (0.6)	18.4

NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Table 3 presents achievement-level results as reported in Braswell et al. (2001), along with projected MB2 results for that same sample. Based on the equipercentile procedure for defining cut points that was described earlier, students who scored at least 34 percent of the possible points on the MB2 form were considered to be performing at or above Basic, those who scored at least 61 percent were considered to be performing at or above Proficient, and those who scored 83 percent and above were considered to be performing at the Advanced level. For the total group, the percentages of students at or exceeding each cut point on MB2 are identical to the corresponding percentages for main NAEP by design. For the subgroups, the procedure used did not guarantee close agreement between results based on the NAEP scale-score metric and those arising from the projected results. However, the projected results agreed closely with the published NAEP results for all the major student groups.

Table 3. Reported achievement-level percentages for the main NAEP sample and projected true-score achievement-level percentages for the main NAEP sample on form MB2

			Main NAEP				MB2			
	Number of students	Percent of students	At *Advanced*	At or above *Proficient*	At or above *Basic*	Below *Basic*	At *Advanced*	At or above *Proficient*	At or above *Basic*	Below *Basic*
Total	13,511	100.0 (0.0)	2.6 (0.3)	26.0 (1.1)	68.7 (1.1)	31.3 (1.1)	2.6 (0.3)	26.0 (1.1)	68.7 (1.1)	31.3 (1.1)
Gender										
Male	6,680	50.7 (0.7)	3.4 (0.4)	28.3 (1.2)	69.6 (1.1)	30.4 (1.1)	3.3 (0.4)	28.2 (1.2)	69.8 (1.1)	30.2 (1.1)
Female	6,831	49.3 (0.7)	1.8 (0.3)	23.6 (1.2)	67.7 (1.2)	32.3 (1.2)	1.9 (0.3)	23.8 (1.2)	67.5 (1.3)	32.5 (1.3)
Race/Ethnicity										
White	8,581	65.9 (0.3)	3.4 (0.4)	33.6 (1.4)	79.6 (1.1)	20.4 (1.1)	3.4 (0.4)	33.7 (1.4)	79.7 (1.1)	20.3 (1.1)
Black	1,795	14.2 (0.2)	0.2 (***)	5.2 (0.9)	38.5 (2.5)	61.5 (2.5)	0.1 (***)	5.1 (0.9)	38.1 (2.5)	61.9 (2.5)
Hispanic	2,239	15.0 (0.3)	0.6 (0.2)	10.3 (1.3)	48.2 (2.1)	51.8 (2.1)	0.7 (0.3)	10.4 (1.3)	47.9 (2.1)	52.1 (2.1)
Type of school										
Public	7,070	89.0 (0.5)	2.4 (0.3)	24.8 (1.2)	66.9 (1.2)	33.1 (1.2)	2.4 (0.3)	24.8 (1.2)	66.9 (1.2)	33.1 (1.2)
Nonpublic	6,441	11.0 (0.5)	5.2 (0.7)	37.8 (1.9)	82.5 (1.6)	17.5 (1.6)	3.9 (0.5)	35.9 (1.1)	82.8 (1.1)	17.2 (1.1)
School lunch eligibility										
Eligible	3,353	32.2 (1.0)	0.3 (0.1)	8.7 (0.8)	45.7 (1.5)	54.3 (1.5)	0.4 (0.1)	8.6 (0.8)	45.4 (1.5)	54.6 (1.5)
Not eligible	10,158	67.8 (1.0)	3.7 (0.5)	33.5 (1.4)	79.4 (1.3)	20.6 (1.3)	3.7 (0.4)	34.3 (1.4)	79.7 (1.2)	20.3 (1.2)

(***) Standard error estimates cannot be accurately determined.
NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Section 3: Comparing Results Obtained with the Market-Basket Short Form to Projected Market-Basket Results

As noted earlier, the notion of implementing market-basket reporting in conjunction with a short-form data collection option had generated some interest at the time of the NAGB redesign of NAEP. Such a configuration could proceed with main NAEP as it is currently designed (i.e., data collected with a matrix-sample design and reported as marginal estimates of reporting-group statistics on a latent-variable scale), while developing market-basket short forms to offer greater flexibility. These short forms could be administered to some subset of the state and national samples during standard NAEP reporting cycles to provide the basis for preliminary fast-track reporting of results. Alternatively, the short forms could be offered as a lower-burden or special lower-cost alternative for states and districts wanting to participate in NAEP during off years.

The data from the current study afforded an opportunity to examine the degree of similarity of results that might be expected when comparing the actual results obtained directly from administering the market-basket form to the projected results derived from the main NAEP sample. More specifically, the samples from the market-basket study that were administered forms MB1 and MB2 and the year-2000 operational sample are randomly equivalent samples from the same target population. Thus, comparing the results obtained by the MB1 and MB2 samples to the projected results of these forms afforded the NAEP program an opportunity to examine the similarities and differences that might be observed in a program configuration in which matrix-sample and short-form data collection were commingled.

As mentioned above, results from the main NAEP samples could be projected as market-basket scores in one of two ways: as market-basket *true scores* or *observed scores*. True scores are more directly analogous to the current NAEP scale-score results, and represent the kind of projection used to produce the prototype market-basket report and the type likely to be employed by the program to effect market-basket reporting. However, the use of a true-score reporting metric does complicate the interpretation of comparisons of projected results to results obtained directly by administering the market-basket forms. The former are estimates of true-score results while the latter are observed-

score results. As is well understood from basic measurement theory, distributional statistics such as achievement-level percentages and standard deviations are not necessarily the same for true scores and observed scores, and such predictable discrepancies can complicate comparisons of results. An alternative approach is to use IRT methods to carry out what Mislevy (2000) refers to as a *two-stage projection* of observed score distributions from main NAEP results (see, e.g., Lord, 1980). The use of a projected observed score would place main NAEP market-basket results and results obtained directly with the market-basket forms on the same scales. However, in that case, projected observed-score results for main NAEP would not mirror the main NAEP scale-score results exactly, as these two sets of results would be on slightly different scales.

In this section of the report, the results obtained by direct administration of the market-basket forms are compared to main NAEP results projected as true scores and observed scores. The first part of this section compares the direct results to the true-score projections obtained using the methods described in the previous section. The second part of this section describes the methods used to carry out observed-score projections of main NAEP results and compares these to the direct results. The final part of this section compares the two sets of projected main NAEP results.

Comparing Actual Market-Basket Results to Projected Market-Basket True-Score Results

Tables 4a and 4b show market-basket estimates of average scores, their standard errors, and standard deviations. These results are shown for the population as a whole and for major NAEP reporting groups. Table 4a shows results for MB2 (the form used for the prototype result), while table 4b shows the results for MB1. In each table, two sets of results are shown: the actual results obtained by directly administering the indicated market-basket form to the market-basket study sample, and the projected true-score results obtained from the main NAEP operational sample. In terms of the overall pattern of results (i.e., differences between reporting groups), agreement is quite good between actual and projected results for MB2 (table 4a). However, there are at least two modest but consistent differences between the two sets of results. The first difference is that the actual standard deviations are slightly, but consistently, larger, both overall and for the

reporting groups, than the projected standard deviations. While this pattern of results could be due to real, but random, differences in the dispersions of proficiency in two samples from which the results are derived, this pattern is also consistent with basic measurement-theory considerations. Classical test theory (Lord & Novick, 1968) predicts that the variance of observed scores (true scores plus error) will be larger than variance of true scores, and this appears to be the case for MB2.

The second small, but consistent, difference is that the projected results are, on average, about a point higher than the actual results. This one point is the equivalent of one-twentieth of a standard deviation, so the difference is small but still worth noting. Again, one possibility is that this discrepancy is due to real differences between the two samples in average proficiency levels, due to random sampling fluctuations. However, there is another possible explanation. The simple scoring algorithm used to produce the MB2 scores assigns a score of 0 to "omitted" items as well as "not-reached" items. The IRT projection procedures have no explicit component to deal with student omissions. The implementation of IRT methods in the current context and in most practical testing applications involves adopting some convention for dealing with student omission, and these conventions are often implicitly in conflict with the simple scoring algorithms used to produce individual student test scores in everyday practice. Although beyond the scope of this project, the effect of treating not-reached items as incorrect responses could be evaluated by rescoring the market-basket forms and treating not-reached items as not reached, meaning that the score would be based solely on the items that had been attempted or omitted within a series of attempted items.

For example, in the current study the procedures implemented for projection made assumptions that are not consistent with the simple scoring algorithm used to produce MB2 results. In the NAEP operational analysis (i.e., the analysis that produced the operational PVs), items that were classified as omitted were treated as incorrect, and items that were classified as not-reached were treated as not presented. In other words, only the omissions embedded within a student response string were scored as incorrect. Omissions occurring outside a student's string of consecutive responses did not contribute one way or the other to estimation of the proficiency distribution. This practice, which is standard in NAEP, is based on research by Mislevy and Wu (1988)

which portrayed it as reasonable and desirable in the context of matrix sample assessments. In the estimation of MB2 true scores conditioned on the estimated latent-variable distribution, an expected item score was generated for all MB2 items. Thus, the projection produced a true-score distribution that assumed students had attempted all the items presented to them. The conventions for treating missing data during the projection stages appear, on their face, to be less punitive and this may be a partial explanation of why the projection appears to overestimate the actual performance obtained when the market basket was directly administered.

Table 4a. Actual observed percent-correct statistics for the market-basket study sample on form MB2, and projected true-score percent-correct statistics for the main NAEP sample on form MB2

	Actual observed results				Projected true-score results			
	Number of students	Percent of students	Mean	Standard deviation	Number of students	Percent of students	Mean	Standard deviation
Total	2,033	100.0 (0.0)	45.6 (0.6)	19.6	13,511	100.0 (0.0)	46.8 (0.5)	19.1
Gender								
Male	1,001	50.4 (1.3)	46.5 (0.8)	19.6	6,680	50.7 (0.7)	47.8 (0.6)	19.6
Female	1,032	49.6 (1.3)	44.6 (0.7)	19.6	6,831	49.3 (0.7)	45.9 (0.5)	18.5
Race/Ethnicity								
White	1,108	66.0 (0.4)	50.1 (0.8)	19.1	8,581	65.9 (0.3)	51.8 (0.6)	18.2
Black	366	14.0 (0.3)	32.4 (1.2)	15.2	1,795	14.2 (0.2)	33.1 (0.7)	14.0
Hispanic	419	15.1 (0.3)	36.9 (0.9)	16.5	2,239	15.0 (0.3)	37.2 (0.8)	16.4
Type of school								
Public	1,800	88.6 (1.1)	44.8 (0.6)	19.4	7,070	89.0 (0.5)	46.0 (0.6)	19.1
Nonpublic	233	11.4 (1.1)	51.2 (2.4)	20.0	6,441	11.0 (0.5)	53.4 (0.5)	17.8
School lunch eligibility								
Eligible	789	33.2 (1.4)	35.8 (0.7)	16.5	3,353	32.2 (1.0)	36.0 (0.5)	15.6
Not eligible	1,244	66.8 (1.4)	50.4 (0.7)	19.2	10,158	67.8 (1.0)	52.0 (0.6)	18.4

NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Table 4b presents actual and projected results for MB1. The same patterns of results are again evident—somewhat larger estimated standard deviations for the actual results than for the projected results and slightly higher average scores overall and for most reporting groups. There is an additional interesting discrepancy between projected and actual results by gender. In the sample to which MB1 was actually administered, there was essentially no difference in performance between males and females. In the main NAEP sample, with results derived using the full NAEP matrix sample instrument, a projected gender difference was found in MB1. This gender difference is evident in the stage-one estimates of the latent variable distributions for these groups, and is carried through to the true-score projections. The projected true-score results for MB1 showed no differences from the actual observed results for females, but the projected true score for male students was approximately 3 points higher than the actual observed score, less than one-sixth of a standard deviation.

It is interesting to note that the MB1 and MB2 projection estimates of the gender difference for the NAEP main sample are virtually identical (1.9 and 2 points, for MB2 and MB1, respectively). This suggests that the absence of a gender difference in the direct results of the MB1 sample is not some strange artifact of the psychometric characteristics of these two test forms (e.g., MB1 being a less discriminating test). Rather, it may be because the evidence about the test performance of males and females provided by the MB1 sample with that instrument is different from the evidence obtained from the MB2 sample or the main NAEP assessment. From the data in this study, it cannot be determined whether the disparities are due to the characteristics of the particular sample that took MB1 or to the specific content of MB1.

Table 4b. Actual observed percent-correct statistics for the market-basket study sample on form MB1 and projected true-score percent-correct statistics for the main NAEP sample on form MB1

	Actual observed results				Projected true-score results			
	Number of students	Percent of students	Mean	Standard deviation	Number of students	Percent of students	Mean	Standard deviation
Total	1,976	100.0 (0.0)	47.1 (0.5)	19.5	13,511	100.0 (0.0)	48.8 (0.5)	18.9
Gender								
Male	973	50.6 (1.3)	46.6 (0.8)	19.6	6,680	50.7 (0.7)	49.8* (0.6)	19.5
Female	1,003	49.4 (1.3)	47.6 (0.8)	19.5	6,831	49.3 (0.7)	47.8 **(0.5)**	18.2
Race/Ethnicity								
White	1,103	66.0 (0.4)	51.8 (0.6)	18.6	8,581	65.9 (0.3)	53.7 **(0.6)**	17.8
Black	359	14.3 (0.3)	34.2 (0.9)	16.0	1,795	14.2 (0.2)	35.2 **(0.8)**	14.5
Hispanic	408	15.0 (0.3)	37.3 (1.0)	16.6	2,239	15.0 (0.3)	39.3 **(0.8)**	16.7
Type of school								
Public	1,746	88.9 (1.1)	46.1 (0.6)	19.4	7,070	89.0 (0.5)	48.0 **(0.6)**	18.9
Non-public	230	11.1 (1.1)	54.5 (1.7)	18.9	6,441	11.0 (0.5)	55.5 **(0.5)**	17.5
School lunch eligibility								
Eligible	775	34.1 (1.5)	35.5 (0.8)	16.1	3,353	32.2 (1.0)	38.1* (0.5)	16.0
Not eligible	1,201	65.9 (1.5)	53.0 (0.7)	18.4	10,158	67.8 (1.0)	54.0 **(0.6)**	18.0

* Statistically significantly different from observed results.

NOTE: The standard errors of the statistics in the table appear in parentheses.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Tables 5a and 5b also show actual versus projected true-score results, but this time the focus is on the percentage of students scoring at each of the achievement levels. Table 5a focuses on form MB2, and shows few differences between the percentage of students in the main sample who were projected to perform at or above each achievement level and the percentage of students in the market-basket sample who actually performed at or above each achievement level. The only statistically significant difference was found for Hispanic students. The percentage of Hispanic fourth-graders scoring at or above *Basic* was about 7 percentage points higher than was projected. Conversely, approximately 7 percent fewer Hispanic students actually performed in the below *Basic* category than was projected.[7]

Table 5b provides actual versus projected achievement-level distributions for form MB1. In terms of the percentage distribution of students performing in each achievement level, there was one consistent difference between the projected and actual percentage of male students scoring at or above each level. A greater percentage of males was projected to score at or above *Basic*, at or above *Proficient*, or at *Advanced* than was actually observed to score at these levels (by approximately 5, 6, and 2 percentage points, respectively). Consistent with these results, about 5 percent more males actually scored below *Basic* on MB2 than were expected to score below *Basic*. The only other statistically significant differences between the actual observed results and projected true-score results were that both for White students and students who were eligible for free or reduced-price school lunch, fewer students than projected scored at or above *Proficient*.

[7] As part of the quality control procedures, analyses for NAEP results include an examination of consistency of projected "percent correct scores" to actual "percent" correct scores, averaged over the individual test booklets in the assessment. The differences between projected and actual results shown in table 5a are quite consistent with the levels of agreement found across booklets in main NAEP as part of the quality control procedures.

Table 5a. Actual observed achievement-level percentages for the market-basket study sample on form MB2, and projected true-score achievement-level percentages for the main NAEP sample on form MB2

	Observed achievement-level distribution						Projected achievement-level distribution					
	Number of students	Percent of students	At *Advanced*	At or above *Proficient*	At or above *Basic*	Below *Basic*	Number of students	Percent of students	At *Advanced*	At or above *Proficient*	At or above *Basic*	Below *Basic*
Total	2,033	100.0 (0.0)	3.0 (0.5)	23.1 (1.2)	69.7 (1.2)	30.3 (1.2)	13,511	100.0 (0.0)	2.6 (0.3)	26.0 (1.1)	68.7 (1.1)	31.3 (1.1)
Gender												
Male	1,001	50.4 (1.3)	2.9 (0.7)	24.7 (1.7)	72.2 (1.8)	27.8 (1.8)	6,680	50.7 (0.7)	3.3 (0.4)	28.2 (1.2)	69.8 (1.1)	30.2 (1.1)
Female	1,032	49.6 (1.3)	3.1 (0.8)	21.5 (1.7)	67.1 (1.6)	32.9 (1.6)	6,831	49.3 (0.7)	1.9 (0.3)	23.8 (1.2)	67.5 (1.3)	32.5 (1.3)
Race/Ethnicity												
White	1,108	66.0 (0.4)	4.1 (0.7)	29.8 (1.6)	78.8 (1.5)	21.2 (1.5)	8,581	65.9 (0.3)	3.4 (0.4)	33.7 (1.4)	79.7 (1.1)	20.3 (1.1)
Black	366	14.0 (0.3)	# (***)	5.8 (1.4)	40.9 (3.8)	59.1 (3.8)	1,795	14.2 (0.2)	0.2 (***)	5.1 (0.9)	38.1 (2.5)	61.9 (2.5)
Hispanic	419	15.1 (0.3)	1.0 (0.2)	7.6 (1.5)	55.1 (2.8)	44.9 (2.8)	2,239	15.0 (0.3)	0.7 (0.3)	10.4 (1.3)	47.9* (2.1)	52.1* (2.1)
Type of school												
Public	1,800	88.6 (1.1)	2.7 (0.6)	22.2 (1.3)	68.3 (1.3)	31.7 (1.3)	7,070	89.0 (0.5)	2.5 (0.3)	24.8 (1.2)	66.9 (1.2)	33.1 (1.2)
Nonpublic	233	11.4 (1.1)	5.1 (1.9)	30.9 (4.7)	80.2 (3.5)	19.8 (3.5)	6,441	11.0 (0.5)	3.9 (0.5)	35.9 (1.1)	82.8 (1.1)	17.2 (1.1)
School lunch eligibility												
Eligible	789	33.2 (1.4)	0.5 (0.3)	7.9 (1.3)	50.6 (2.2)	49.4 (2.2)	3,353	32.2 (1.0)	0.4 (0.1)	8.6 (0.8)	45.4 (1.5)	54.6 (1.5)
Not eligible	1,244	66.8 (1.4)	4.2 (0.7)	30.7 (1.6)	79.2 (1.3)	20.8 (1.3)	10,158	67.8 (1.0)	3.7 (0.4)	34.3 (1.4)	79.7 (1.2)	20.3 (1.2)

The estimate rounds to zero.
* Statistically significantly different from observed results.
(***) Sample size is insufficient to permit a reliable estimate.
NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Table 5b: Actual observed achievement-level percentages for the market-basket study sample on form MB1, and projected true-score achievement-level percentages for the main NAEP sample on form MB1

	Observed achievement-level distribution						Projected achievement-level distribution					
	Number of students	Percent of students	At *Advanced*	At or above *Proficient*	At or above *Basic*	Below *Basic*	Number of students	Percent of students	At *Advanced*	At or above *Proficient*	At or above *Basic*	Below *Basic*
Total	1,976	100.0 (0.0)	2.3 (0.4)	22.7 (1.2)	67.0 (1.3)	33.0 (1.3)	13,511	100.0 (0.0)	2.6 (0.3)	26.0 (1.0)	68.7 (1.1)	31.3 (1.1)
Gender												
Male	973	50.6 (1.3)	1.9 (0.4)	22.7 (1.8)	64.9 (2.0)	35.1 (2.0)	6,680	50.7 (0.7)	3.4* (0.4)	28.4* (1.1)	69.9* (1.2)	30.1* (1.2)
Female	1,003	49.4 (1.3)	2.7 (0.7)	22.8 (1.8)	69.1 (1.7)	30.9 (1.7)	6,831	49.3 (0.7)	1.8 (0.3)	23.6 (1.2)	67.4 (1.3)	32.6 (1.3)
Race/Ethnicity												
White	1,103	66.0 (0.4)	2.9 (0.5)	29.4 (1.6)	76.6 (1.5)	23.4 (1.5)	8,581	65.9 (0.3)	3.4 **(0.4)**	33.5* (1.4)	79.7 **(1.2)**	20.3 **(1.2)**
Black	359	14.3 (0.3)	0.3 (***)	5.2 (1.4)	40.4 (2.6)	59.6 (2.6)	1,795	14.2 (0.2)	0.2 **(***)**	5.3 **(0.4)**	38.1 **(2.5)**	61.9 **(2.5)**
Hispanic	408	15.0 (0.3)	0.4 (***)	8.1 (1.6)	48.1 (2.5)	51.9 (2.5)	2,239	15.0 (0.3)	0.6 **(0.2)**	10.6 **(1.4)**	47.9 **(2.1)**	52.1 **(2.1)**
Type of school												
Public	1,746	88.9 (1.1)	2.1 (0.4)	21.5 (1.4)	65.6 (1.4)	34.4 (1.4)	7,070	89.0 (0.5)	2.4* (0.3)	24.7* **(1.1)**	66.9 **(1.2)**	33.1 **(1.2)**
Nonpublic	230	11.1 (1.1)	3.8 (1.6)	32.3 (4.3)	78.5 (3.4)	21.5 (3.4)	6,441	11.0 (0.5)	4.2 **(0.4)**	36.2 **(1.2)**	83.0 **(1.0)**	17.0 **(1.0)**
School lunch eligibility												
Eligible	775	34.1 (1.5)	0.1 (***)	6.5 (0.9)	43.0 (2.1)	57.0 (2.1)	3,353	32.2 (1.0)	0.4* (0.1)	8.9 (0.8)	45.6 (1.6)	54.4 (1.6)
Not eligible	1,201	65.9 (1.5)	3.4 (0.5)	31.1 (1.7)	79.4 (1.3)	20.6 (1.3)	10,158	67.8 (1.0)	3.7 (0.4)	34.1 (1.3)	79.6 (1.2)	20.4 (1.2)

* Statistically significantly different from observed results.
(***) Sample size is insufficient to permit a reliable estimate.
NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Comparing Actual Market-Basket Results to Projected Market-Basket Observed-Score Results

This section describes the methods used to carry out observed-score projections of main NAEP results, and then compares these results to the actual results.

Converting NAEP Scale Scores to a Market-Basket Observed-Score Metric

Converting existing NAEP scale-score PVs to the observed-score metric was a bit more complicated than converting to the true-score metric, but was still relatively straightforward. For this study, *Monte Carlo* procedures were used to obtain the estimates. Specifically, a *plausible item response* was generated for each item and examinee, based on the probabilities of a correct response, given the estimated item parameters for the market-basket forms and main NAEP PVs. These simulated item responses, or item scores, were then summed to get a *plausible market-basket observed score*. Finally, this plausible market-basket score was converted into a percent-correct metric.

Tables 6a and 6b show estimated means, standard errors, and standard deviations for MB2 and MB1, respectively. Actual results based on the market-basket study samples and projected observed-score results based on the main NAEP samples are shown. The findings from these comparisons show some similarities to the findings from comparisons of actual results with projected true-score results, as well as some differences. For example, once again there appears to a consistent pattern of slightly higher performance in the projected results than was evident in the results obtained by directly administering the market basket forms. As noted for the earlier comparisons, this pattern could be indicative of real differences between the samples or could be an artifact of the analysis approaches used to deal with items that students omitted. In addition, projected observed-score results and actual results for MB1 again appear to show a different pattern of gender differences. One clear point of difference between the earlier set of comparisons and these, however, involves the estimated standard deviations. The projected observed-score standard deviations appear somewhat larger than the actual standard deviations. This result may again be due to the differences in the treatment of students' omitted responses.

Table 6a. Actual observed percent-correct statistics for the market-basket study sample on form MB2 and projected observed-score percent-correct statistics for the main NAEP sample on form MB2

	Actual observed results				Projected observed-score results			
	Number of students	Percent of students	Mean	Standard deviation	Number of students	Percent of students	Mean	Standard deviation
Total	2,033	100.0 (0.0)	45.6 (0.6)	19.6	13,511	100.0 (0.0)	46.8 (0.5)	20.1
Gender								
Male	1,001	50.4 (1.3)	46.5 (0.8)	19.6	6,680	50.7 (0.7)	47.6 (0.6)	20.5
Female	1,032	49.6 (1.3)	44.6 (0.7)	19.6	6,831	49.3 (0.7)	46.0 (0.5)	19.6
Race/Ethnicity								
White	1,108	66.0 (0.4)	50.1 (0.8)	19.1	8,581	65.9 (0.3)	51.7 (0.6)	19.2
Black	366	14.0 (0.3)	32.4 (1.2)	15.2	1,795	14.2 (0.2)	33.2 (0.8)	15.5
Hispanic	419	15.1 (0.3)	36.9 (0.9)	16.5	2,239	15.0 (0.3)	37.3 (0.8)	17.9
Type of school								
Public	1,800	88.6 (1.1)	44.8 (0.6)	19.4	7,070	89.0 (0.5)	46.0 (0.6)	20.1
Nonpublic	233	11.4 (1.1)	51.2 (2.4)	20.0	6,441	11.0 (0.5)	53.2 (0.5)	19.0
School lunch eligibility								
Eligible	789	33.2 (1.4)	35.8 (0.7)	16.5	3,353	32.2 (1.0)	35.9 (0.5)	16.8
Not eligible	1,244	66.8 (1.4)	50.4 (0.7)	19.2	10,158	67.8 (1.0)	52.0 (0.6)	19.5

NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Table 6b. Actual observed percent-correct statistics for the market-basket study sample on form MB1, and projected observed-score percent-correct statistics for the main NAEP sample on form MB1

	Actual observed results				Projected observed-score results			
	Number of students	Percent of students	Mean	Standard deviation	Number of students	Percent of students	Mean	Standard deviation
Total	1,976	100.0 (0.0)	47.1 (0.5)	19.5	13,511	100.0 (0.0)	48.9* (0.5)	20.0
Gender								
Male	973	50.6 (1.3)	46.6 (0.8)	19.6	6,680	50.7 (0.7)	49.9* (0.6)	20.5
Female	1,003	49.4 (1.3)	47.6 (0.8)	19.5	6,831	49.3 (0.7)	47.8* (0.6)	19.5
Race/Ethnicity								
White	1,103	66.0 (0.4)	51.8 (0.6)	18.6	8,581	65.9 (0.3)	53.6* (0.6)	19.1
Black	359	14.3 (0.3)	34.2 (0.9)	16.0	1,795	14.2 (0.2)	35.4 **(0.8)**	15.8
Hispanic	408	15.0 (0.3)	37.3 (1.0)	16.6	2,239	15.0 (0.3)	39.4 **(0.9)**	18.0
Type of school								
Public	1,746	88.9 (1.1)	46.1 (0.6)	19.4	7,070	89.0 (0.5)	48.0* (0.6)	20.0
Nonpublic	230	11.1 (1.1)	54.5 (1.7)	18.9	6,441	11.0 (0.5)	55.4 **(0.5)**	18.8
School lunch eligibility								
Eligible	775	34.1 (1.5)	35.5 (0.8)	16.1	3,353	32.2 (1.0)	38.3* (0.5)	17.4
Not eligible	1,201	65.9 (1.5)	53.0 (0.7)	18.4	10,158	67.8 (1.0)	53.9* (0.6)	19.2

* Statistically significantly different from observed results.
NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Comparing Projected Results Obtained Using the True-Score and Observed-Score Metrics

Table 7 presents the two sets of projected market-basket results from the main NAEP sample side by side. A comparison of the two sets of projected results shows patterns directly predictable from basic measurement theory. Estimates of average observed scores and true scores, overall and by subgroup, are nearly identical. This is to be expected because, under classical measurement theory, expected observed scores are equal to their true-score counterparts. Estimated true-score standard deviations are slightly smaller than their observed-score equivalents, both for the national results and for each of the reporting groups. Such a result is again predicted by classical test theory results, in which the variance of observed scores is the sum of the variance of true scores plus the variance of measurement errors.

Table 7. Projected true-score and projected observed-score percent-correct statistics for the main NAEP sample on forms MB1 and MB2

	MB1				MB2			
	Projected true score		Projected observed score		Projected true score		Projected observed score	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Total	48.8 (0.5)	18.9	48.9 (0.5)	20.0	46.8 (0.5)	19.1	46.8 (0.5)	20.1
Gender								
Male	49.8 (0.6)	19.5	49.9 (0.6)	20.5	47.8 (0.6)	19.6	47.6 (0.6)	20.5
Female	47.8 (0.5)	18.2	47.8 (0.6)	19.5	45.9 (0.5)	18.5	46.0 (0.5)	19.6
Race/Ethnicity								
White	53.7 (0.6)	17.8	53.6 (0.6)	19.1	51.8 (0.6)	18.2	51.7 (0.6)	19.2
Black	35.2 (0.8)	14.5	35.4 (0.8)	15.8	33.1 (0.7)	14.0	33.2 (0.8)	15.5
Hispanic	39.3 (0.8)	16.7	39.4 (0.9)	18.0	37.2 (0.8)	16.4	37.3 (0.8)	17.9
Type of school								
Public	48.0 (0.6)	18.9	48.0 (0.6)	20.0	46.0 (0.6)	19.1	46.0 (0.6)	20.1
Nonpublic	55.5 (0.5)	17.5	55.4 (0.5)	18.8	53.4 (0.5)	17.8	53.2 (0.5)	19.0
School lunch eligibility								
Eligible	38.1 (0.5)	16.0	38.3 (0.5)	17.4	36.0 (0.5)	15.6	35.9 (0.5)	16.8
Not eligible	54.0 (0.6)	18.0	53.9 (0.6)	19.2	52.0 (0.6)	18.4	52.0 (0.6)	19.5

NOTE: The standard errors of the statistics in the table appear in parentheses.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000 Mathematics Market Basket.

Concluding Remarks

More information about the technical, methodological, and policy issues needs to be developed in order to implement market-basket reporting and/or short-form data gathering in NAEP. A key part of that information will be a specific model of NAEP reports that uses market-basket reporting.

Although the main study described here focused on the issues of market-basket reporting from a large administration rather than a short-form administration and on the differences between true-score and observed-score reporting, other psychometric issues were explored using the data from this study. Specifically, context effects of embedded market-basket items were examined to determine whether the results of embedded items would differ markedly from the results of the same items assembled in a single form. In addition, it was important to examine whether the differences between MB1 and MB2 were due to differences in the forms or the samples. These two issues are discussed here.

The 2000 design required a substantial investment of resources to identify a supplemental sample of 8,000 students to whom intact market-basket and linking forms could be directly administered. It is reasonable to ask why a more efficient design that embedded the market-basket items or blocks into the main NAEP could not be used. For example, at one extreme, the results for MB1 theoretically could be estimated directly from the main NAEP sample, since all its items appear somewhere among the 13 main NAEP blocks. So, it would seem efficient to add the MB2 items, or blocks, to the main NAEP booklet design to obtain the results.

There are several reasons why an *embedded-item* or *embedded-block* design was not employed. One reason was the need to check the adequacy of IRT projection techniques to accurately reproduce the distributions of directly obtained observed scores. However, the principal reason for avoiding an embedded-item approach was concern over context effects. While context effects may be deemed non-problematic in some applications, in others they can introduce serious difficulties into the accurate measurement of trends. (For more information on context effects, see Brennan, 1992; Harris, 2003; and Leary & Dorans, 1985).

The year-2000 market-basket study provided data to examine the extent and impact of context effects in estimating large-population statistics. For example, two distinct sets of item-parameter estimates existed for the items in MB1. One set consisted of the estimates for these items as they appeared in the main NAEP sample. The second set consisted of estimates obtained for these items as they appeared in intact blocks within the market-basket sample. In a separate study, the two sets of item parameter were used to obtain two sets of item-, block- and book-characteristics curves, and a few items showed nonoverlapping characteristics curves or a mild presence of context effect (Tay-Lim, Wang, & Mazzeo, 2001). However, when the items were aggregated to block and book level, the two sets of characteristics curves overlapped, showing minimal context effect at the block or book level. Further analysis in this area would be helpful to determine whether the use of embedded items would result in projected results that are sufficiently comparable to actual results to be of use operationally.

For example, for a given sample of students (e.g., the combined MB1, MB2, LINK1, and LINK2 sample) two sets of projections (true-score metric and observed-score metric projections) of average scores and PACs could be obtained. One set of projections would use the main NAEP item-parameter estimates while the other would use the market-basket item-parameter estimates. Comparing the two sets of results would give the program additional findings on the magnitude of context effects and provide guidance in the design of future market-basket calibration studies. However this comparison has not yet been conducted and, therefore, is not discussed in this technical report.

As with most research endeavors, there will be limits on the generalizability of what has been learned to other situations in NAEP, particularly to other subject areas. In many respects, the mathematics assessment may represent an ideal subject for market-basket reporting and short-form data gathering. The mathematics assessment makes considerable use of multiple-choice and short constructed-response items, and most items are discrete (i.e., not organized in large sets that are associated with a single set of stimulus materials or reading passages.). Because of the nature of the existing NAEP mathematics framework, there is some reason for optimism that short forms of reasonable length (displaying both sufficient content coverage to illustrate the full item pool and

sufficient measurement precision to support the use of short forms as a data-gathering strategy) could be constructed. In other content areas, the challenges may be somewhat greater.

Consider the NAEP reading assessment as a counterexample. The current framework calls for the measurement of reading for three distinct purposes (for literary experience, for information, and to perform a task) at grades 8 and 12. The framework also calls for the use of intact, authentic reading passages of substantial length and the use of multiple-choice, short constructed-response, and extended constructed-response item formats. Each 25-minute block of the reading assessment consists of one reading passage and a set of related questions. In NAEP's current matrix sample design, student testing time is held to 50 minutes (i.e., two 25-minute blocks). As a result, no eighth- or twelfth-grade student is measured in all three reading areas.

A reading market basket for reporting purposes most certainly would need to contain at least one block of items for each purpose of reading. Thus, it would involve defining a "synthetic" market-basket test form that is greater in length than any one form in the main NAEP assessment, as currently administered. Use of a market-basket short form as a data collection device would pose some interesting complications. It may be possible that a relatively small number of 50-minute "partial market-basket test forms" could be administered and observed score results from these forms used to projected results for a "synthetic" market basket. Such an approach, however, would almost certainly invoke some form of complicated statistical machinery. If a version of short-form data gathering that assumes simple analysis procedures (i.e., procedures that do not require IRT or direct estimation of reporting-group statistics) was to be considered, the reading market-basket short forms would likely require at least three blocks of items (i.e., 75 minutes of student testing time). Moreover, it is arguable whether one passage from each reading purpose would be considered sufficient to adequately exemplify the framework, or whether results based on short forms of this length would be sufficiently form independent to be used as an alternative short-form assessment system. A six-block reading market-basket short-form test (i.e., 150 minutes of student testing time) might be more in line with what is required. However, from the perspective of a student or school, a test of this length would clearly not be viewed as a "short form" of NAEP.

In summary, much was learned from the 2000 study about the issues and challenges involved with implementing a market-basket system in NAEP. The two market test forms created for this study were intended to be parallel. They produced similar, though not identical results, when administered to randomly equivalent national samples of students. The two forms differed slightly in overall difficulty level, and the relative difficulty of the two test forms was not constant across all subgroups studied. Based on the data from this study it is not possible to determine whether these differences in test performance are related to the specific content of the forms or to sampling fluctuations.

Two different approaches to converting NAEP results to market-basket scores were described and illustrated. One-stage projection involved the conversion of NAEP latent-variable results to a market-basket true-score scale. One-stage-projection results were used in a prototype market-basket report developed by the program and described in section 2 of the report. Two-stage projection involved the conversion of NAEP latent variable results to a market-basket observed-score scale. Results for a nationally-representative sample of students from both sets of projections were compared with each other and with the results actually obtained by directly administering the market basket to separate nationally-representative samples. While the two kinds of projection results were generally similar, differences consistent with what one would expect from basic measurement theory considerations were evident between them. Furthermore, both sets of projection results were similar, in most cases, to actual results obtained by directly administering the market baskets to separate, randomly equivalent samples. There were, however, some notable differences. Of particular note was the general pattern of projected results being slightly higher than actual results. This pattern is consistent with the fact that, for the actual results, not all students attempted all the items in the market basket and all such items were scored as incorrect. In contrast, the projected results assume all items attempted by all students. This underscores the important fact that standard IRT methods for projecting results onto different forms are based on the assumption of power tests, not speeded tests, and can be expected to work less well for tests with increasing amounts of omissions and not-reached items.

All things considered, it may be more prudent ultimately for the NAEP program to consider implementing market-basket reporting when new frameworks and trend lines are introduced. Under these conditions, issues associated with reconciling the new market-basket results with results previously reported in other metrics would not arise, and item development work, field-testing plans, and the design of the assessment instruments could be established from "the ground up" to appropriately support the system. Separate decisions can be made on a subject-by-subject basis as to whether offering a short-form data-gathering option makes sense (given the nature of the content area and its associated framework), and whether the necessary development and field-test work would be economically feasible.

References

Allen, N. L., Carlson, J. E., & Donoghue, J. R. (2001). Overview of Part I: The design and implementation of the 1998 NAEP. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 technical report* (NCES 2001-509, pp. 5–23). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Allen, N. L., Carlson, J. E., Johnson, E. G., & Mislevy, R. J. (2001). Scaling procedures. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 technical report* (NCES 2001-509, pp. 227–246). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17,* 95–110.

Bock, R. D. (1996). *Domain-referenced reporting in large-scale educational assessments.* Commissioned paper to the National Academy of Education, for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., & Santapau, S. L. (2001). *The Nation's Report Card: Mathematics 2000* (NCES 2001-517). U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education, 5,* 225–264.

DeVito, P. J., & Koening, J. A. (Eds.) (2000). *Designing a market basket for NAEP: Summary of a workshop.* Committee on NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting. Board on Testing and Assessment. National Research Council. Washington, DC: National Academy Press.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996, July 1). *Design/feasibility team report to the National Assessment Governing Board.* Retrieved November 1, 2005 from www.nagb.org/pubs/appj.html.

Harris, D. J. (2003, April). A *conceptual synthesis of context effects*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Hecht, L., & Swineford, F. (1981). *Item analysis at Educational Testing Service.* Princeton, NJ: Educational Testing Service.

Horkay, N. (Ed.). (1999). *The NAEP Guide* (NCES 2000-456). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Jenkins, F. J., Chang, H. H., & Kulick, E. M. (1999). Data analysis for the mathematics assessment. In N. L. Allen, J. E. Carlson, & C. A. Zelenak (Eds.), *The NAEP 1996 technical report* (NCES 1999-452, pp. 255–290). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55,* 387–413.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mazzeo, J., & Lazer, S., & Zieky, M. J. (In press). Monitoring educational progress with group score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

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

Mislevy, R. J. (1998). Implications of market basket reporting for achievement-level setting. *Applied Measurement in Education, 11,* 49–84.

Mislevy, R. J. (2000, February). *Evidentiary relationships among data gathering methods and reporting scales in surveys of educational achievement.* Paper presented to the National Academy of Sciences Committee on NAEP Reporting Practices, Washington, DC.

Mislevy, R. J. & Wu, P. K. (1988) *Inferring examinee ability when some item responses are missing* (ETS Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (1995). *Developing student performance levels for the National Assessment of Educational Progress.* Policy statement.

National Assessment Governing Board. (1996). *Redesigning the National Assessment of Educational Progress.* Policy statement.

Tay-Lim, B. S.-H., Wang, M. M., & Mazzeo, J. (2001, April). *The effects of contexts in the item performance of NAEP market-basket forms.* Paper presented at meeting of the National Council on Measurement in Education, Seattle, WA.

Tay-Lim, B. S.-H., Wang, M. M., & Kulick, E. (2002, April). *An empirical evaluation of various methods for linking market-basket forms to main NAEP*. Paper presented at meeting of the National Council on Measurement in Education, New Orleans, LA.