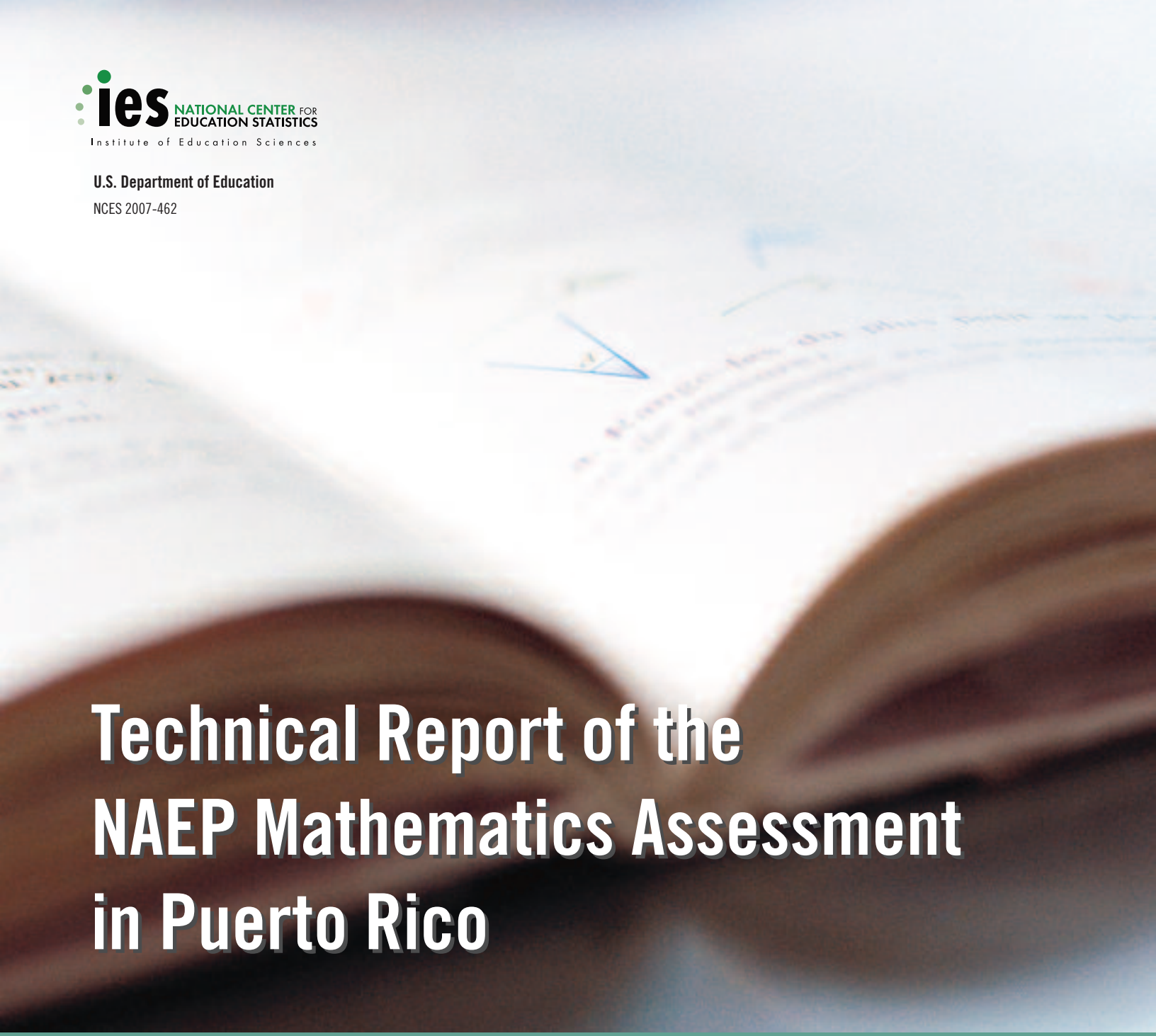
ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

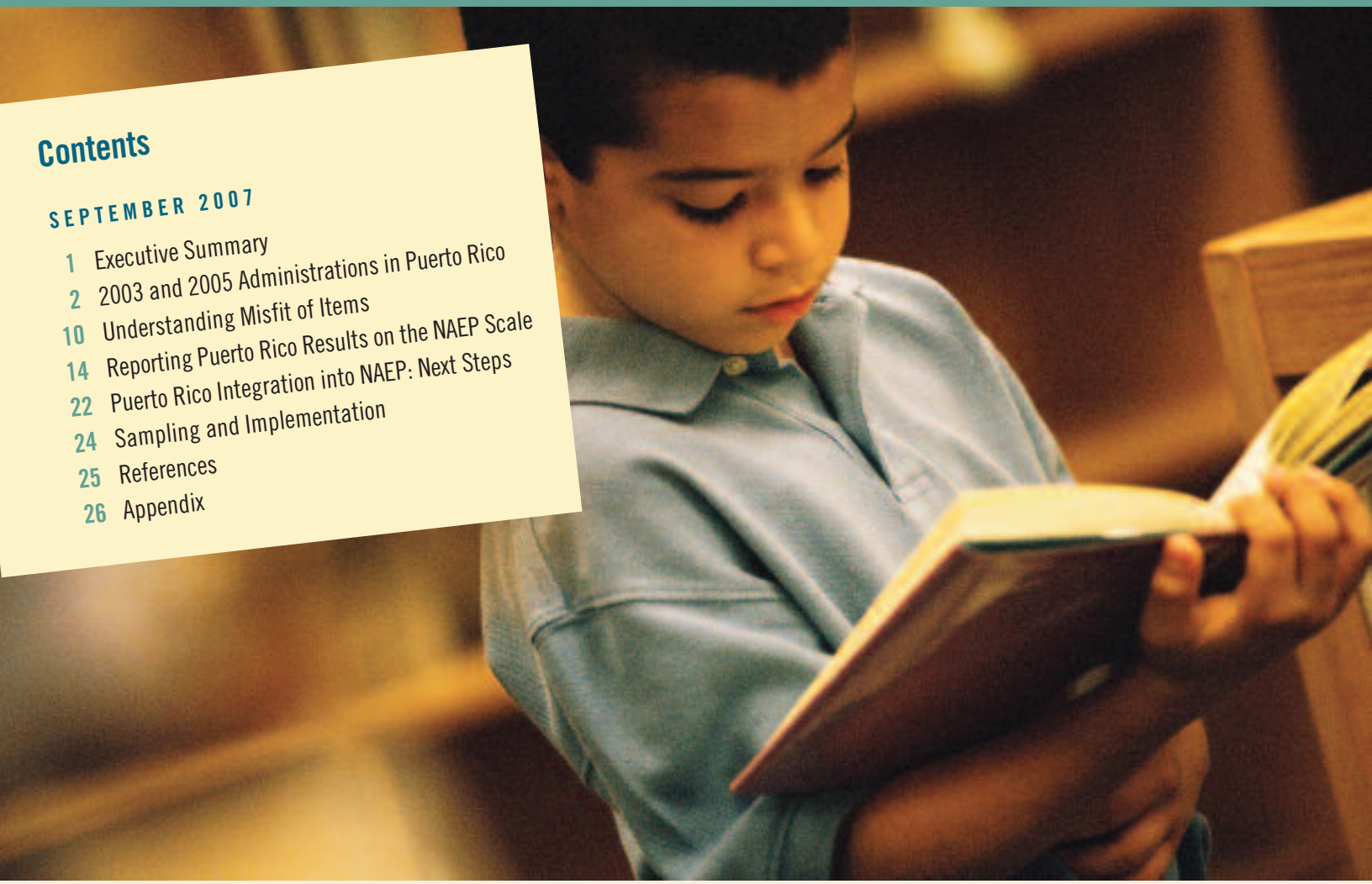
U.S. Department of Education
NCES 2007-462



Technical Report of the NAEP Mathematics Assessment in Puerto Rico

FOCUS ON STATISTICAL ISSUES

National Assessment of Educational Progress



Contents

SEPTEMBER 2007

1 Executive Summary
2 2003 and 2005 Administrations in Puerto Rico
10 Understanding Misfit of Items
14 Reporting Puerto Rico Results on the NAEP Scale
22 Puerto Rico Integration into NAEP: Next Steps
24 Sampling and Implementation
25 References
26 Appendix

What is The Nation's Report Card™?

The Nation's Report Card™ informs the public about the academic achievement of elementary and secondary students in the United States and its jurisdictions, including Puerto Rico. Report cards communicate the findings of the National Assessment of Educational Progress (NAEP), a continuing and nationally representative measure of achievement in various subjects over time. The Nation's Report Card™ compares performance among states, urban districts, public and private schools, and student demographic groups.

For over three decades, NAEP assessments have been conducted periodically in reading, mathematics, science, writing, history, geography, and other subjects. By making objective information available on student performance at the national, state, and local levels, NAEP is an integral part of our nation's evaluation of the condition and progress of education. Only information related to academic achievement and relevant variables is collected. The privacy of individual students is protected, and the identities of participating schools are not released. NAEP is a congressionally mandated project of the National Center for Education Statistics (NCES) within the Institute of Education Sciences of the U.S. Department of Education. The Commissioner of Education Statistics is responsible for carrying out the NAEP project. The National Assessment Governing Board oversees and sets policy for NAEP.

Executive Summary

In 2003, a trial NAEP mathematics assessment was administered in Spanish to public school students at grades 4 and 8 in Puerto Rico. Based on preliminary analyses of the 2003 data, changes were made in administration and translation procedures for the 2005 NAEP administration in Puerto Rico. This report describes the content and administration of the trial NAEP mathematics assessments in Puerto Rico in 2003 and 2005, problems with item misfit in the 2003 data, results of a special validity analysis, and plans to integrate Puerto Rico into the national sample in future administrations.

The 2003 trial NAEP mathematics assessment in Puerto Rico was administered in Spanish. Preliminary analysis of the 2003 Puerto Rico data raised concerns that the items were not functioning as they did in other jurisdictions. In Puerto Rico, there were larger amounts of missing data, fewer correct responses than expected for every content area, and a more frequent mismatch between expected and actual student performance on items (item misfit) compared to other jurisdictions.

To improve data quality, modifications to translation and administration procedures were made for the 2005 assessment in Puerto Rico. These changes included revision of the administration script, provision of an additional 10 minutes for each of the two timed sections of the assessment, and enhanced translation procedures. Analysis of the 2005 data showed fewer missing responses and a higher percentage of correct responses compared to 2003. Despite these improvements in data quality, there was concern about the validity of reporting Puerto Rico results on the NAEP scale.

To address this concern, a validity study was conducted using the 2003 and 2005 results in Puerto Rico. The analysis involved eliminating items that exhibited misfit in the Puerto Rico sample, thus yielding a restricted scale. Performance on the reduced set of items (restricted scale) was then compared to the corresponding performance on the full set of items (full scale) for the nation and Puerto Rico. The two scales agreed to within three-tenths of a scale point in the nation, but there was a 2- to 3-point difference between the scales in Puerto Rico. However, this difference does not change the ranking of Puerto Rico among other participating jurisdictions. These findings indicate that Puerto Rico results could be reported on the NAEP 0-500 scale.

All jurisdictions receiving federal Title 1 funds, including Puerto Rico, are required to participate in NAEP in fourth-and eighth-grade every other year beginning in 2003. The 2003 and 2005 NAEP administrations in Puerto Rico were considered trials and results were not reported with those of other jurisdictions. In future NAEP administrations, the intent is to include Puerto Rico as part of the national sample. For the 2007 administration, NCES increased the involvement of Puerto Rico educators in the development and translation review process of the NAEP mathematics assessment. Steps are in place to move Puerto Rico toward full integration into the NAEP sampling, data collection, and reporting for the 2009 administration.

About this report

This report is one of a series of three on the administration and results of the 2003 and 2005 trial NAEP mathematics assessments in Puerto Rico available at http://nationsreportcard.gov/puertorico_2005/. The first report, *Mathematics 2003 and 2005 Performance in Puerto Rico: Highlights,* presents results for Puerto Rico and the nation in terms of NAEP scale scores and achievement levels. The second report, *Mathematics 2005 Performance in Puerto Rico: Focus on the Content Areas,* provides results by content area and includes a discussion of student performance on a sample of items. This, the third report, focuses on the technical considerations of the trial assessments and plans to include Puerto Rico as part of the national sample in future administrations.

Chapter 1

2003 and 2005 Administrations in Puerto Rico

The NAEP mathematics assessment was translated into Spanish to allow Puerto Rico to participate on a trial basis in the 2003 administration. In 2005, the NAEP mathematics assessment was again administered to fourth- and eighth-grade public school students in Puerto Rico. A primary goal for the trial administrations was to report Puerto Rico results on the national scale. This chapter describes the 2003 and 2005 NAEP mathematics administrations in Puerto Rico, the data quality concerns that emerged in 2003, and the changes made in 2005 to address those concerns.

2003 NAEP Mathematics Assessment in Puerto Rico

Title 1 of the Elementary and Secondary Education Act of 1965, as amended, requires all jurisdictions receiving federal Title 1 funds to participate in NAEP at fourth- and eighth-grade in reading and mathematics every other year beginning in 2003. The U.S. Department of Education decided that Puerto Rico should not participate in the NAEP reading, because that assessment measures a student's ability to read in English, and Spanish is the language of instruction in Puerto Rico. The 2003 NAEP mathematics assessment was translated into Spanish to allow Puerto Rico to participate on a trial basis.

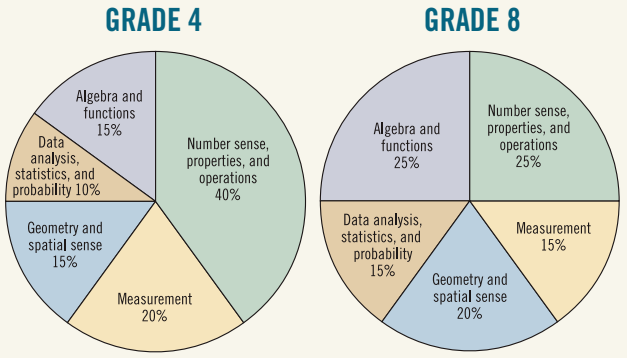
The primary goal of the 2003 trial administration in Puerto Rico was to administer the NAEP mathematics assesment using the same procedures used in other jurisdictions because, in the future, the results of the Puerto Rico assessment are to be compared over time and with those of other jurisdictions. The NAEP administration in Puerto Rico—content, item types,

sampling and administration, and scoring procedures—was consistent with that of other jurisdictions.

Content. The content of the mathematics assessment is based on a framework that describes in detail how mathematics should be assessed by NAEP. The NAEP mathematics framework specifies the content to be assessed at each grade level and the percentage of questions to be assessed in each of five content areas. In 2003, the five content areas were (1) number sense, properties, and operations; (2) measurement; (3) geometry and spatial sense; (4) data analysis, statistics, and probability; and (5) algebra and functions. The percentage distribution of items for grades 4 and 8 are shown in figure 1.1.

Figure 1.1

Percentage distribution of NAEP mathematics items, by grade and content area in 2003



SOURCE: Mathematics Framework for the 2003 National Assessment of Educational Progress, National Assessment Governing Board, U.S. Department of Education, September 2002.

In addition to content, the framework specified that each item should measure one of three mathematical abilities: conceptual understanding, procedural knowledge, and problem solving. The frameworks are available at http://www.nagb.org/pubs/pubs.html.

Item types. The NAEP mathematics assessment includes a combination of multiple-choice and constructed-response items. Multiple-choice items require students to select the correct response from four or five possible choices. Responses to these items are scored correct or incorrect. Short constructed-response items require students to provide answers to computation problems or to describe solutions in one or two sentences. Extended constructed-response items require students to provide written answers of more than a sentence or two. These items are designed to measure students' abilities to reason, communicate, and make connections between concepts and skills, either across the five mathematics content areas or from mathematics to other curricular areas. Responses to constructed-response items are scored correct or incorrect, or they are scored with one or more levels of partial credit.

Translation. A subset of the NAEP mathematics materials was first translated for the 1996 English/Spanish bilingual accommodation by a team of bilingual test development specialists. Mathematics textbooks used in bilingual education were consulted to ensure plausible contexts and accurate mathematical terminology. In 2002, as preparations began for the 2003 administration in Puerto Rico, mathematics items that had been translated into Spanish for bilingual accommodation were evaluated for use in a Spanish-only version of the assessment to be administered in Puerto Rico.

A panel of eight Spanish-speaking educators, including three Puerto Ricans teaching in the United States, evaluated the initial translation of the items. The goal was to produce a psychometrically equivalent assessment rather than a word-by-word translation from English to Spanish.[1] As part of the panel review, adaptation and translation parameters that accounted for linguistic and cultural considerations particular to Puerto Rico were established to ensure that vocabulary selected during the translation process would be appropriate for students in grades 4 and 8. These parameters were then applied in the translation of the 2003 NAEP mathematics assessment for administration in Puerto Rico.

[1] A discussion of the issues involved in conducting assessments in multiple languages can be found in Hambleton, R.K., Merenda, P. F., and Spielberger, C.D. (2005).

Following the panel reviews, bilingual language specialists completed editorial and equity/fairness reviews on all questions.

Sampling and administration. For the trial NAEP administration in 2003, approximately 100 public schools and 3,000 students at each of grades 4 and 8 were sampled in Puerto Rico. Details of the sampling procedures and participations rates can be found on page 24 of this report.

Consistent with procedures used in other jurisdictions, the administration period began the last week of January and continued through the first week of March 2003. Each student received a booklet that contained two 25-minute blocks of items, for a total assessment length of 50 minutes. In addition to the mathematics items, students were asked to respond to questions about their educational background, including questions specific to their mathematics instruction and experiences.

The assessment sessions were conducted by administrators who were not staff members of participating schools. These administrators were hired and trained in Puerto Rico to minimize the impact of linguistic and cultural differences on the administration. Administrators' materials, including manuals and directions, were translated into Spanish.

Scoring. NAEP mathematics results are reported in two ways: as average scores on the NAEP mathematics scale and as percentages of students attaining different NAEP mathematics achievement levels. Results are reported for performance overall and for performance in each of the five content areas on a 0–500 scale. Scale scores are computed for groups of students, not for individual students.

In addition to scale scores, results are presented in terms of mathematics achievement levels as adopted by the Governing Board. Achievement levels are intended to measure how well students' actual achievement matches the achievement expected of them. For each grade tested, the Governing Board has adopted three achievement levels: *Basic, Proficient,* and *Advanced*. Information on how the Governing Board sets achievement levels can be found at http://nces.ed.gov/nationsreportcard//pubs/main1996/97951.asp.

NAEP achievement levels

The three NAEP achievement levels from lowest to highest are *Basic, Proficient,* and *Advanced.*

Basic: This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Proficient: This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.

Advanced: This level signifies superior performance.

Cut scores

Cut scores represent the minimum score required for performance at each NAEP achievement level. The mathematics cut scores on the 0–500 NAEP scale that define the lower boundaries of each of the achievement levels are

	Grade 4	Grade 8
Basic	214	262
Proficient	249	299
Advanced	282	333

Preliminary Analysis of the 2003 Administration

The 2003 NAEP in Puerto Rico was the first attempt to conduct an entire administration in a language other than English. Preliminary analysis of the data for the 2003 mathematics assessment indicated three important differences between Puerto Rico and other jurisdictions. In Puerto Rico, there was a higher percentage of missing responses to items, higher percentage of incorrect responses to items, and higher levels of item misfit compared to other jurisdictions.

High percentage of missing data. Table 1.1 shows the percentage of missing responses for Puerto Rico and the nation in 2003. Missing responses include omitted and not reached items. An item was considered omitted if a student skipped that question but answered one or more questions following it. An item was considered as not reached when neither that question nor any question following it in the section was answered. For the 2003 assessment overall, each question was not answered, on average, by 25 percent of fourth-graders in Puerto Rico, compared to 7 percent in the nation. For eighth-graders in 2003, the respective percentages of missing responses were 19 percent in Puerto Rico and 5 percent in the nation. In every content area in both grades, there were more missing responses in Puerto Rico than the nation.

The percentage of missing responses in Puerto Rico was higher than the nation for all item types in both grades. For the short constructed-response items, on average, 39 percent of the responses were missing on these items at grade 4, and 29 percent of the responses at grade 8 were missing for these items. For the nation, the corresponding percentages of missing responses to short constructed-response items were 9 percent and 6 percent at grades 4 and 8, respectively.

Table 1.1

Average percentage of missing responses on NAEP mathematics assessment for public school students in Puerto Rico and the nation at grades 4 and 8, by content area and item type in 2003

	Grade 4			Grade 8		
		Percent missing responses			Percent missing responses	
Item characteristic	Number of items	Puerto Rico	Nation	Number of items	Puerto Rico	Nation
Overall	**179**	**25***	**7**	**195**	**19***	**5**
Content area						
Number sense, properties, and operations	75	23*	6	51	14*	4
Measurement	32	21*	6	30	16*	4
Geometry and spatial sense	27	25*	6	36	22*	5
Data analysis, statistics, and probability	19	28*	5	29	24*	7
Algebra and functions	26	32*	10	49	20*	6
Item type						
Multiple choice	114	16*	4	126	11*	4
Short constructed response	57	39*	9	60	29*	6
Extended constructed response	8	59*	22	9	62*	25

* Puerto Rico significantly different ($p < .05$) from the nation.
NOTE: Missing responses include omitted and not reached items.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Mathematics Assessment.

Higher percentage of incorrect responses. In addition to the high percentage of missing responses, there was also a higher percentage of incorrect responses in Puerto Rico compared to the nation. This finding may indicate a problem with how items functioned in Puerto Rico, or it may reflect differences between the knowledge and skills of students in Puerto Rico and those of students in the nation. Table 1.2 presents the average of the mean item scores overall, by mathematics content area, and by item type. For multiple-choice and short constructed-response items that are scored dichotomously (correct or incorrect), the mean item score reflects the proportion of correct responses for a particular item.

All of the extended constructed-response items and some of the short constructed-response items are scored using multilevel scoring guides. Students may be credited with a partially correct response. For items such as these, the mean item score is defined as the average proportion of maximum score points received. The data in table 1.2 present the average across all items.

Overall, the average of the mean item score for Puerto Rico was 0.27 at grades 4 and 8 compared to 0.54 for the nation. For each of the five content areas of the NAEP mathematics assessment in both grades 4 and 8, the average of the mean item scores was lower in Puerto Rico than in the nation. The same pattern of results was found for each item type.

High levels of item misfit. The most problematic result to emerge from the 2003 Puerto Rico mathematics assessment was the relatively high incidence of item misfit. Item misfit is defined as a mismatch between expected and actual student performance on an item. Although items functioned as expected in the nation, the translated items did not function as expected in Puerto Rico. Chapter 2 provides a discussion of item misfit.

Table 1.2

Difference between average of the mean item scores on NAEP mathematics assessment for public school students in Puerto Rico and the nation at grades 4 and 8, by content area and item type in 2003

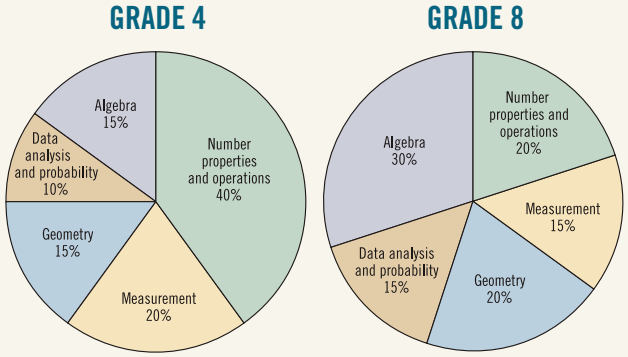
Item characteristic	Grade 4			Grade 8		
	Puerto Rico	Nation	Difference	Puerto Rico	Nation	Difference
Overall	**.27***	**.54**	**.27**	**.27***	**.54**	**.27**
Content area						
Number sense, properties, and operations	.26*	.55	.29	.31*	.59	.28
Measurement	.30*	.52	.22	.25*	.50	.25
Geometry and spatial sense	.32*	.53	.21	.29*	.55	.26
Data analysis, statistics, and probability	.24*	.59	.35	.22*	.50	.28
Algebra and functions	.20*	.50	.31	.26*	.55	.29
Item type						
Multiple choice	.34*	.58	.24	.30*	.57	.27
Short constructed response	.15*	.50	.34	.24*	.54	.31
Extended constructed response	.05*	.29	.24	.06*	.25	.19

* Puerto Rico significantly different ($p < .05$) from the nation.
NOTE: Details may not sum to total due to rounding. Differences are based on unrounded estimates.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Mathematics Assessment.

2005 NAEP Mathematics Assessment in Puerto Rico

The NAEP mathematics assessment was administered on a trial basis for a second time in Puerto Rico in 2005. Details of the sampling procedures and participation rates can be found on page 24 of this report. As in 2003, approximately 100 public schools and 3,000 students at each of grades 4 and 8 were sampled. In both the nation and Puerto Rico, there was a change to the content of the assessment to reflect changing curricular emphases and objectives specified in the NAEP mathematics framework. In 2005, names of some of the content areas changed, but the percentage of questions to be assessed in each of five content areas remained the same for grade 4 (figure 1.2). At grade 8, there was a 5 percent decreased emphasis on number properties and operations and a corresponding 5 percent increased emphasis on algebra compared to 2003.

Figure 1.2
Distribution of NAEP mathematics items, by grade and content area in 2005



SOURCE: Mathematics Framework for the 2005 National Assessment of Educational Progress, National Assessment Governing Board, U.S. Department of Education, September 2004.

In addition to specifying content, the 2005 framework calls for an assessment that measures different levels of mathematical complexity. Each level of mathematical complexity (high, moderate, and low) includes aspects of knowing and doing mathematics, such as reasoning, performing procedures, understanding concepts, or solving problems. The level of complexity of a question is determined by the cognitive demands that it places on students. For example, a question with a high level of complexity at grade 4 might ask students to explain and justify their solutions to a problem.

Changes to the 2005 Puerto Rico Administration

Concerns about the quality of the 2003 Puerto Rico data led to three changes for the 2005 trial NAEP in Puerto Rico. These changes included enhanced translation procedures, revised administration procedures, and the addition of workshops to develop an understanding of NAEP and to encourage participation in NAEP.

Enhanced translation procedures. An independent translation verification review was conducted by a language translation company hired by the National Center for Education Statistics (NCES). These translation review procedures were conducted after the 2003 assessment booklets had been printed, and therefore the changes resulting from this process were first implemented for the 2005 assessment. Three other enhancements were made to the translation procedures. First, two Puerto Rico mathematics teachers were added to the expert review panel. These two teachers at grades 4 and 8 in Puerto Rico, provided valuable perspective on how students in Puerto Rico might interpret specific wording and respond to certain contexts. Second, some contexts and language used in the mathematics items were adapted to the unique linguistic and cultural characteristics of Puerto Rico. Third, the full set of background questionnaires was translated and adapted for use in Puerto Rico using the same parameters that guided the translation of the assessment items.

Revised administration procedures. Two changes were implemented to the administration process. First, the administration script in Puerto Rico was revised to give students explicit directions to move on to the next question rather than continue with an item for which they did not know the answer. The revised script also included extensive explanations of the different item types in the assessment. These revisions were made

because students in Puerto Rico may not have been familiar with the NAEP assessment format.

Second, in an attempt to reduce the prevalence of missing responses, students were provided with an additional 10 minutes to complete each of the two timed mathematics sections. In other participating jurisdictions, students are given 50 minutes to complete the assessment and students receiving accommodations are given an additional 10 minutes per section for a total of 70 minutes. Thus, in 2005, students in Puerto Rico were given the same amount of time to complete the assessment as were students receiving accommodations in other jurisdictions.

Motivational workshops. Two motivational workshops were held in San Juan and Ponce a few weeks prior to the 2005 NAEP administration. The purpose of the workshops was to increase interest in NAEP and to help administrators learn how to better encourage teachers and students to participate in the 2005 NAEP administration. About 150 principals and other administrators from the schools selected for the 2005 assessment participated. Workshop sessions provided information about the content, purpose, and reporting goals of NAEP. Participants were given sample questions and manipulatives that had been used on past assessments, and together they explored strategies for helping students do well on assessments such as NAEP. The workshops were conducted by Spanish-speaking NAEP assessment developers and attended by representatives from the Puerto Rico Department of Education.

Preliminary Analysis of the 2005 Administration

Analysis of the 2005 data from Puerto Rico indicated a decrease in the percentage of missing responses and an increase in the percentage of correct responses compared to 2003. These changes may be due to modifications in the translation and administration procedures described above or to increased student performance in grades 4 and 8, or both.

Table 1.3 shows the reduction in the percentage of missing responses between 2003 and 2005 at grades

Table 1.3

Difference between average of the percentage missing responses in 2003 and 2005 on NAEP mathematics assessments for public school students in Puerto Rico at grades 4 and 8, by content area and item type

	Grade 4			Grade 8		
	Percent missing responses			Percent missing responses		
Item characteristic	2003	2005	Difference	2003	2005	Difference
Overall[1]	**25***	**11**	**14**	**18***	**9**	**9**
Content area						
Number sense, properties, and operations	24*	10	13	15*	8	7
Measurement	22*	9	13	17*	8	8
Geometry and spatial sense	25*	13	12	25*	13	11
Data analysis, statistics, and probability	30*	14	16	20*	11	9
Algebra and functions	26*	12	14	16*	6	9
Item type						
Multiple choice	16*	6	10	11*	4	7
Short constructed response	37*	18	19	28*	17	11
Extended constructed response	66*	37	29	59*	30	29

* 2003 significantly different ($p < .05$) from 2005.

[1] The NAEP mathematics framework used in 2003 differs from the framework used in 2005. The table lists the content areas for 2003. In 2005, the content areas were number properties and operations, measurement, geometry, data analysis and probability, and algebra.

NOTE: Average percentage of missing responses is calculated using those items that were administered in both 2003 and 2005. Missing responses include omitted and not reached items. Differences are based on unrounded estimates.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

4 and 8. Items that were common across the two assessment years (2003 and 2005) were used in this analysis. For each content area and for each item type and grade, the average percentage of missing responses in 2005 was lower than in 2003. Nevertheless, the amount of missing responses remains higher in Puerto Rico than in the nation (table 1.3).

Table 1.4 shows the average of the mean item scores in 2003 and 2005 by grade and item type. There were significant improvements at grades 4 and 8 overall and for some of the content areas. With the exception of grade 4 multiple-choice items, the average of the mean item score increased from 2003 to 2005 for all item types at each grade level. Despite these improvements, large mean item differences remain between Puerto Rico and the nation.

The changes made between the 2003 and 2005 administrations were intended to address concerns about the quality of the 2003 data. Results of the 2005 administration showed higher mean item scores and lower percentages of missing data compared to 2003. However, assessment scores in Puerto Rico continued to lag significantly behind the rest of the nation. Further, in Puerto Rico there are more missing responses to items and lower mean item scores compared to the nation. Details of the performance of students in Puerto Rico can be found in *Mathematics 2003 and 2005 Performance in Puerto Rico: Highlights* available at http://nationsreportcard.gov/puertorico_2005/. As described in chapter 3, additional analyses were undertaken to address concerns about the extent to which translated items used in Puerto Rico did not function as expected.

Table 1.4

Difference between the average of the mean item scores on 2003 and 2005 NAEP mathematics assessments administered to public school students in Puerto Rico, by content area and item type

	Grade 4			Grade 8		
	Mean score			Mean score		
Item characteristic	2003	2005	Difference	2003	2005	Difference
Overall[1]	**.27***	**.28**	**.02**	**.27***	**.28**	**.02**
Content area						
Number sense, properties, and operations	.27*	.29	.02	.28*	.30	.02
Measurement	.28	.28	.01	.21	.22	.01
Geometry and spatial sense	.32*	.35	.03	.29*	.31	.02
Data analysis, statistics, and probability	.23	.23	#	.23*	.26	.02
Algebra and functions	.23*	.25	.02	.29*	.31	.02
Item type						
Multiple choice	.33	.34	.01	.28*	.30	.01
Short constructed response	.16*	.20	.04	.25*	.27	.02
Extended constructed response	.03*	.07	.04	.11*	.14	.03

Rounds to zero.

* 2003 significantly different ($p < .05$) from 2005.

NOTE: Average of the mean item scores is calculated using those items that were administered in both 2003 and 2005. Differences are based on unrounded estimates.

[1] The NAEP mathematics framework used in 2003 was revised in 2005. The table lists the content areas for 2003. In 2005, the content areas were number properties and operations, measurement, geometry, data analysis and probability, and algebra.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Chapter 2

Understanding Misfit of Items

The NAEP mathematics assessment consists of a set of items designed to measure a variety of ways of knowing and doing mathematics. Each student responds to a subset of the total items in each of the five content areas. Responses are scored, and performance is reported for groups of students not for individuals. Using item response theory (IRT) models, performance can be predicted on any given item. The relationship between predicted and actual performance is fundamental to producing the NAEP scale for reporting results. Preliminary analyses of the Puerto Rico 2003 results showed a discrepancy between actual and predicted performance for a large number of items (item misfit). This chapter provides a discussion of item misfit and its implications for interpreting the Puerto Rico results.

Predicting Student Performance

The NAEP scales are designed to measure proficiency with respect to a framework of knowledge and skills. The assessment items are examples of the knowledge and skills represented in the framework. Like many other assessment programs, NAEP releases some items and replaces them with new ones in the next assessment. Therefore, NAEP requires mechanisms for reporting performance on the same scale even though the items change over time.

Statistical models based on Item Response Theory (IRT) are used to estimate the distribution of student proficiency on the NAEP scale and link different items to a common scale (Lord 1980; Hambleton, Swaminathan, and Rogers 1991; Embretson and Reise 2000). IRT models characterize each item on an assessment. For any given item, the probability of a correct response (0.0 to 1.0) can be plotted against a continuum of proficiency levels.

Typically, the resulting item characteristic curve (ICC) shows that as proficiency or ability level increases, the probability of a correct response increases as well.[2] More proficient students (higher mathematical ability) are more likely than less proficient students to answer an item correctly.

The form of the ICC is determined by item difficulty and item discrimination. The more difficult an item, the lower the probability of a correct response by students with low mathematics knowledge and skill (low ability or proficiency). Item discrimination is reflected in the steepness of the ICC. Thus, the steeper the curve (slope), the better an item can discriminate between those with more or less proficiency. These characteristics of the ICC are useful in examining the correspondence between expected and actual performance on a set of items.

Item Misfit

An ICC establishes the expected performance of students. The extent to which students actually perform as expected on an item is referred to as item fit. Item fit is evaluated by comparing the proportion of examinees within a relatively narrow proficiency range who respond correctly to an item with the expected performance of examinees in that range. Discrepancy between expected and actual performance is defined as item misfit.

In Puerto Rico, some items exhibited item misfit. To illustrate the concept of item misfit, it is helpful to first consider the ICC for a perfectly fitting item (figure 2.1). In the figure, each circle represents the actual performance of a set of examinees. Those at the low end of the proficiency scale have a low probability of a correct item response. As the curve shifts to the right, students with a greater proficiency have a higher probability of correctly answering this item. In this figure, actual performance (circles) falls directly on the ICC.

[2] Theta is used to denote ability or proficiency on the underlying construct the assessment is attempting to measure/estimate. The measure of ability has no inherently meaningful units of measurement. The most common, albeit arbitrary, choice of the theta scale is one in which the mean is zero and the standard deviation is 1, the z score scale.

Figure 2.2 illustrates a discrepancy between actual performance and expected performance. Circles above the line show groups of students who performed better than expected, and circles below the line show students who did not perform as well as expected on this item. As noted above, for a perfectly fitting item, most of the less proficient students are expected to get the item wrong, and most of the more proficient students are expected to get the item right. In figure 2.2, the probability of a correct response was higher or lower than expected for a given proficiency level.

Figure 2.1
Example of an item characteristic curve of a perfectly fitting item

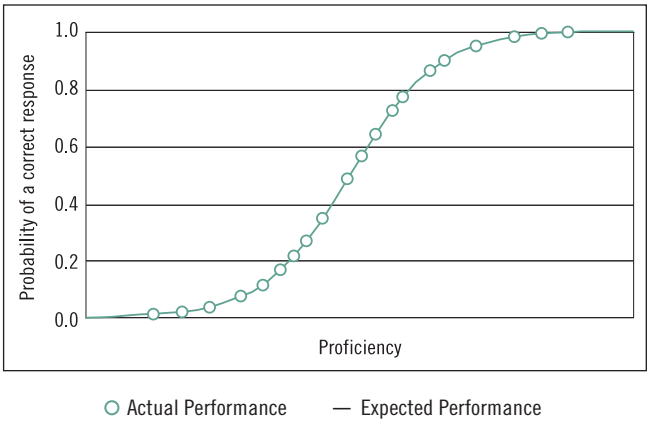
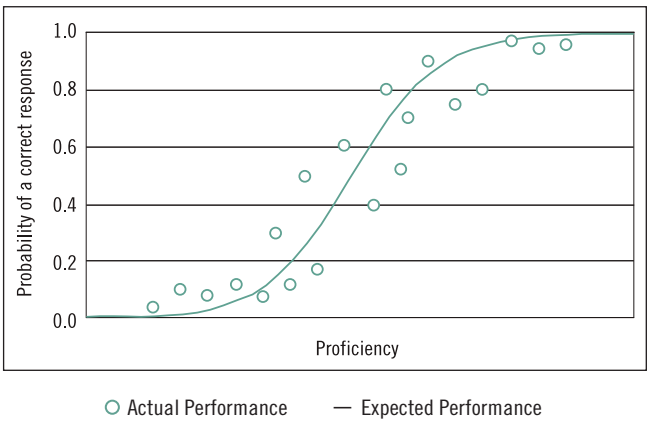


Figure 2.2
Example of an item characteristic curve showing item misfit



Sources of Item Misfit

Expected performance on the 2003 NAEP mathematics items was estimated from the national sample (all participating jurisdictions except Puerto Rico). The high levels of item misfit in Puerto Rico may indicate curricular differences between Puerto Rico and the nation, or translation errors, or an assessment that is too difficult, or some combination of these issues.

• ***Curricular differences.*** Jurisdictions differ in when they teach certain skills or concepts. For example, one jurisdiction may teach multiplication of fractions in grade 4 while another teaches it in grade 5. A student who is proficient in mathematics, but who has not yet been taught multiplication of fractions, would have difficulty solving an item of this type.

• ***Translation.*** In Puerto Rico, NAEP is administered using a Spanish language instrument. For the 2003 assessment, the focus of the translation process was on developing a single Spanish version that could be universally administered to students of Puerto Rican, Cuban, Mexican, Central American, and Spanish ancestry. In subsequent assessments, a translation review process evaluated the appropriateness of the translations for administration in Puerto Rico only. The process of translating the assessment from English to Spanish may result in subtle changes that alter the meaning of an item, or make assessment items easier or more difficult.

• ***Mismatch between assessment difficulty and student proficiency.*** Item difficulty was estimated based on item tryouts with the national sample. Items were not pilot tested in Puerto Rico prior to the 2003 administration. Some of the items may have been at a different level of difficulty for students in Puerto Rico than for students in the nation. A battery of assessment items that is more difficult than the students' levels of proficiency can increase item non-response, guessing, and confound proficiency with speed. Preliminary analysis of the 2003 data showed higher levels of missing responses and more incorrect responses for students in Puerto Rico compared to students in the national sample.

Implications of Item Misfit

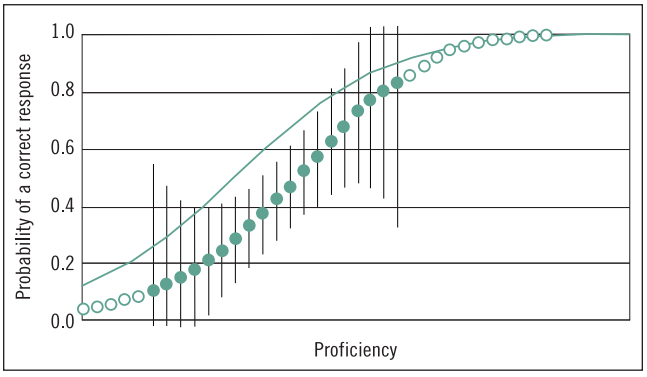
Although many of the NAEP assessment items did not function as expected in Puerto Rico, this does not preclude reporting the Puerto Rico data on the NAEP scale. Item misfit indicates a discrepancy between actual and predicted performance and suggests the need for further analysis to (1) understand the nature of the item misfit, and (2) evaluate the implications of the item misfit for reporting Puerto Rico results on the NAEP scale.

These issues can be addressed by examining the ICCs of items that do not function as expected. The sample graphs 2.3 and 2.4 include standard error bars that display the margin of error around each point. Where the margin of error grows too large to display, the

points are displayed as empty circles indicating that not enough students are at that proficiency range to provide meaningful data.

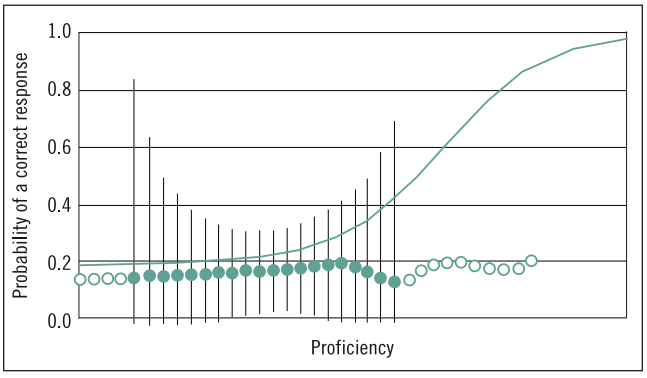
• ***Items too difficult for the population.*** Figure 2.3 shows an example of an item that is too difficult for the examinees. Two things are important to notice here. First, actual student performance (circles) is below the curve of the predicted performance except at the highest levels of proficiency. Second, there are very few students in Puerto Rico who are at the higher end of the proficiency scale. Items such as these do not contribute much to estimates of overall student proficiency because the vast majority of students get the item wrong.

Figure 2.3
Example of an item characteristic curve for an item that is too difficult for students



NOTE: Circles indicate actual student performance. Empty circles indicate insufficient numbers of students at a proficiency level to provide meaningful data.

Figure 2.4
Example of an item characteristic curve of an item that is relatively unrelated to student proficiency



NOTE: Circles indicate actual student performance. Empty circles indicate insufficient numbers of students at a proficiency level to provide meaningful data.

• ***Item performance unrelated to proficiency.*** Figure 2.4 presents an item that is relatively unrelated to student proficiency. Performance on the item is expected to increase with increased proficiency (upward curve at the right of the graph). However, in Puerto Rico, actual student performance (indicated by the circles) is relatively flat. For this item, the probability of getting the answer correct is not related to student proficiency. Students with more mathematics knowledge and skill (high end of the proficiency scale) have the same probability of getting this item correct as do students with less mathematics knowledge and skill (low end of the proficiency scale).

The item may have been difficult for students in Puerto Rico because the concept had not been sufficiently introduced prior to the assessment. Alternatively the item may have been difficult for students in Puerto Rico to understand due to errors in translation from English to Spanish that changed the meaning of the item, or perhaps more than one correct or nearly correct response has been introduced through translation. Such items do not create bias in that they will not lead to systematically higher or lower scores, but they do increase the measurement error (precision with which performance can be estimated).

In sum, many of the NAEP items functioned differently in Puerto Rico than predicted. Item misfit primarily affects aggregate statistics by reducing the precision of the estimates. Item misfit may signal problems with the assessment and potentially biased results. However, the results may be an accurate reflection of the proficiency of the group being assessed. Chapter 3 examines the accuracy of the Puerto Rico results in terms of the NAEP scale.

Chapter 3

Reporting Puerto Rico Results on the NAEP Scale

The Puerto Rico results for 2003 and 2005 were not reported with those of other jurisdictions because of concerns that translated items used in Puerto Rico were not functioning as expected. A set of analyses was conducted to examine whether the NAEP mathematics assessment measured student achievement in Puerto Rico in the same way that it measured student achievement in the nation. This chapter describes the analyses and the implications for reporting Puerto Rico results on the NAEP scale.

In future NAEP administrations, the intent is to include Puerto Rico as part of the national sample. Although Puerto Rico participated in NAEP mathematics assessments in 2003 and 2005, results were not included in the overall national estimates because of concerns about the quality of the Puerto Rico data. In Puerto Rico there were higher amounts of missing data, fewer correct responses, and higher levels of item misfit compared to the nation. The release of the 2003 and 2005 results was delayed pending additional analyses to examine the accuracy of the Puerto Rico results in terms of the NAEP scale.

The analyses proceeded in four steps. First, items identified as problematic for Puerto Rico in the 2003 and 2005 NAEP mathematics assessment were removed from the total set of administered items to produce a reduced set of items. Second, content coverage in terms of the five NAEP mathematics content areas was compared for the full and reduced sets of items. Third, the two sets of items were calibrated and placed on the NAEP scale to allow for comparisons. Fourth, performance on the full set of items was compared to performance on the reduced set of items for Puerto Rico, the nation,

and the jurisdictions comprising the national sample. In conducting these analyses, comparisons are made for public school students because only public school students in Puerto Rico participated in the NAEP mathematics assessments in 2003 and 2005.

Identification of Problematic Items

Items were identified as problematic for one of two reasons: (1) translation errors and differential item functioning, and (2) items that did not function well in Puerto Rico.

Translation errors and differential item functioning. One possible explanation for item misfit is translation errors. Following the preparation of the 2003 test booklets, an independent translation review was undertaken by a language translation company at the request of NCES. The translation verification process classified items as having no translation errors, minor errors, moderate errors, or severe errors. However, even severe translation errors did not account for most of the item misfit in Puerto Rico.

All items used in the NAEP 2003 mathematics assessment underwent an analysis of differential item function (DIF),[3] a technique to identify item misfit. For each DIF analysis, the performance of Puerto Rico (the focal group) was compared to the performance of the following reference groups: the nation, District of Columbia, Virgin Islands, American Samoa, and students in the national public sample who identified themselves as Puerto Rican. The goal was to determine whether items functioned differently in Puerto Rico than in the nation or other jurisdictions for reasons unrelated to ability. Results indicated that items flagged as having significant DIF did not account for most or all of the item misfit.

[3] An item exhibits DIF if the probability of doing well on the item depends on group membership, even after controlling for ability. The DIF methods used in this analysis were the Mantel-Haenszel chi-square procedure for dichotomous items and the Mantel procedure for polytomous items (Holland and Wainer 1993).

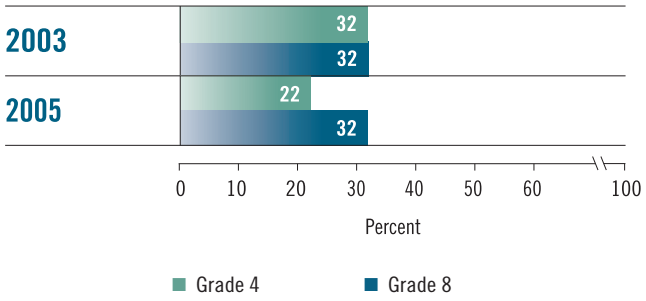
Items from the 2003 NAEP mathematics assessment identified as (1) having severe translation errors, and (2) as displaying significant DIF were considered problematic. The translation errors identified in 2003 were corrected prior to the 2005 NAEP administration in Puerto Rico. Consequently, the translation review and DIF analysis of items with translation errors were not repeated for the 2005 data.

Items that did not function well in Puerto Rico. In 2003 and 2005, the majority of items identified as problematic were items that did not function well for the Puerto Rico sample. These included (1) items which were not correlated with total score for the Puerto Rico sample; (2) items with no variance in response for the Puerto Rico sample; (3) items that did not discriminate between students with more or less ability (flat ICC for Puerto Rico); and (4) items that exhibited misfit between the theoretical and empirical ICCs for the Puerto Rico sample.

Content Coverage of Full and Reduced Item Sets

Problematic items were removed from the full set of items to create a reduced set of items at grades 4 and 8 for 2003 and 2005. On average, across years and grades, 30 percent of items were eliminated from the total set of items (figure 3.1).

Figure 3.1
Percentage of total NAEP mathematics assessment items eliminated, by year and grade



SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Removing items from the total set of items could change the construct being measured or shift the emphasis from one content area to another. Because the reduced set of items is a subset of the full set of items, chi square tests were used to compare the distribution of kept items (items that comprise the reduced set) to the distribution of dropped items (total items minus kept items). As shown in table 3.1, the percentage distribution of items by content area and by mathematical ability for the reduced set of items is not significantly different from the corresponding distribution of items for the dropped set of items used in the NAEP 2003 administration. The percentage of items by item type in the reduced set differed significantly from the distribution by item type in the dropped set. For 2005, the pattern of results was the same at grade 4. At grade 8, the distribution of dropped items by content area, item type, or mathematical ability did not differ significantly from the distribution of kept items in 2003 or 2005. Appendix A includes additional information on the dropped, kept, and full sets of items.

Table 3.1
Percentage distribution of dropped, kept, and full item sets at grade 4 for NAEP mathematics assessment, by item characteristics in 2003

Item characteristic	Dropped	Kept	Full
Content area			
Numbers sense, properties, and operations	32	47	42
Measurement	19	17	18
Geometry and spatial sense	18	14	15
Data analysis, statistics, and probability	11	11	11
Algebra and functions	21	11	15
Item type*			
Multiple choice	79	57	64
Short constructed response	18	39	32
Extended constructed response	4	5	4
Mathematical ability			
Conceptual understanding	46	37	40
Problem solving	40	34	36
Procedural knowledge	14	29	24

* Distribution of dropped items significantly different ($p < .05$) from distribution of kept items.
NOTE: Detail may not sum to totals due to rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Mathematics Assessment.

Table A-1 shows the percentage distribution of dropped, kept, and full sets of items at grade 4, by year and item characteristics. Table A-2 shows the percentage distribution of dropped, kept, and full set of items at grade 8, by year and item characteristics. Table A-3 shows the minimum and maximum mean item scores for the full and reduced sets of items for Puerto Rico and the nation, by grade and year.

Calibration of Full and Restricted Scales

Two scales were created—a full scale and a restricted scale. The full scale was based on the total set of NAEP mathematics items administered in 2003 and 2005. The restricted scale was based on a reduced set of items (those items remaining after excluding the problematic items). For the restricted scale, it was necessary to recalibrate the assessment to account for the changes in the treatment of the test items (exclusion of problematic items) and the inclusion of Puerto Rico students. Consistent with procedures used in the operational NAEP, the five mathematics subscales were calibrated separately and then combined in a linear combination.

Linking constants were calculated to transform the restricted-scale results from 2003 to those from the full-scale in 2003, thus placing them on the NAEP scale and allowing the results to be compared.[4] The 2005 restricted-scale results were then linked to the 2003 restricted-scale results using standard NAEP procedures. The methodology used to equate scores during this linking procedure can introduce error into the estimates of student ability. However, the statistical procedures and methodology needed to properly estimate the impact of the linking error in this context have not been developed for NAEP.

[4] NAEP uses common population linking procedures (Allen, Donoghue, and Schoeps 2001).

Comparison of Full and Restricted Scales

Full- and restricted-scale comparisons were conducted for Puerto Rico, the nation, and individual jurisdictions comprising the national sample. Three questions guided this analysis. First, is there a difference in the overall mean score? Second, do the percentages of students performing at each achievement level vary by scale? Differences in overall mean scores may or may not affect the percentages of students performing at each achievement level. Third, are the results of the full- and restricted-scale comparisons across jurisdictions consistent with the results of these comparisons in Puerto Rico and the nation? Taken together, the results of these analyses provide evidence to support reporting Puerto Rico results on the NAEP scale and thereby allowing comparisons between Puerto Rico and the nation.

Mean scale scores. The mean NAEP mathematics scores using the full and restricted scales were compared for Puerto Rico and for the nation (table 3.2).[5] For Puerto Rico, the difference between the mean score on the full scale and the restricted scale ranges from 1.9 to 3.4 points across grades and NAEP administrations. Mean scores were higher on the restricted scale than on the full scale in Puerto Rico. For the nation, the mean score on the two scales agrees to within three tenths of a scale point at grades 4 and 8 in both 2003 and 2005. The difference between the full- and restricted-scale mean scores was statistically significant in 2005. The statistical significance is most likely due to the large sample sizes.

[5] Differences between scale scores or percentages were calculated using unrounded numbers. In some instances, the result of the subtraction differs from what would be obtained by subtracting the rounded values shown in the accompanying figure or table.

Table 3.2

Mean full- and restricted-scale scores for NAEP mathematics assessment in the nation and Puerto Rico, by year and grade

	Full scale		Restricted scale		Full minus restricted scale	
Jurisdiction	**Mean**	**Standard error**	**Mean**	**Standard error**	**Mean difference**	**Standard error**
Puerto Rico						
2003						
Grade 4	179	1.0	182	1.0	-3.0*	0.13
Grade 8	212	1.0	214	1.1	-2.0*	0.16
2005						
Grade 4	183	0.9	187	0.9	-3.4*	0.13
Grade 8	218	1.0	220	1.1	-1.9*	0.11
Nation						
2003[1]						
Grade 4	234	0.2	234	0.2	0.0	0.02
Grade 8	276	0.3	276	0.3	0.0	0.04
2005						
Grade 4	237	0.2	237	0.2	0.3*	0.02
Grade 8	278	0.2	277	0.2	0.1*	0.03

* Difference statistically significant ($p < .05$).
[1] An artifact of the linking procedure is perfect agreement between the full and restricted scales in 2003.
NOTE: The standard errors do not include linking error and thus may be underestimates of the true standard errors. The statistical procedures and methodology needed to properly estimate the impact of the linking error have not been developed for NAEP.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Achievement levels. NAEP reports performance in terms of three achievement levels: *Basic, Proficient,* and *Advanced*. Because few students performed at the *Advanced* level in Puerto Rico, data are provided for *Basic* and *Proficient* levels only. Details are available in *Mathematics 2003 and 2005 Performance in Puerto Rico: Highlights* available at http://nationsreportcard.gov/puertorico_2005/.

Table 3.3 presents the percentage of students at or above *Basic* and the percentage of students at or above *Proficient*. In Puerto Rico, a higher percentage of students scored at or above *Basic* on the restricted scale than on the full scale. The differences ranged from 1.16 to 1.71 percentage points across grades 4 and 8 in 2003 and 2005. In the nation, fewer students scored at or above the *Proficient* level on the restricted scale than on the full scale. The differences ranged from 0.32 to 0.82 percentage points across grades in 2003 and 2005.

Table 3.3

Percentage of students performing at selected NAEP achievement levels on the full and restricted scale in the nation and Puerto Rico, by grade and year

	Grade 4				Grade 8			
	2003		2005		2003		2005	
Jurisdiction	Full	Restricted	Full	Restricted	Full	Restricted	Full	Restricted
Puerto Rico								
Percent below *Basic*	91*	89	88	87	96*	94	94*	93
Percent at or above *Basic*	9*	11	12*	13	4*	6	6*	7
Percent at or above *Proficient*	#	#	#	#	#	#	#	#
Nation								
Percent below *Basic*	24	24	21	21	33	33	32	32
Percent at or above *Basic*	76	76	79	79	67	67	68	68
Percent at or above *Proficient*	31*	31	35*	35	27*	27	28*	28

Rounds to zero.
* Full scale significantly different ($p < .05$) from restricted scale.
NOTE: Differences between percentages are calculated using unrounded numbers. In some instances, the result of the subtraction differs from what would be obtained by subtracting the rounded numbers shown in the table.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Comparisons across other jurisdictions. The performance of individual jurisdictions that comprise the national sample is summarized in Table 3.4. Row 1 shows the mean difference between the full and restricted scale for the national sample overall. Across jurisdictions comprising the national sample, the maximum and minimum mean scale score differences are shown in rows 2 and 3, respectively. Row 4 shows the percentage of participating jurisdictions with mean score differences of less than 1 scale point.

As shown in Table 3.4, the differences between mean scores on the full and restricted scale range from 0.5 points to 1.6 points across jurisdictions that comprise the national sample. The maximum and minimum differences reported in rows 2 and 3, respectively indicate that some states had higher mean scores on the full scale and some states had higher mean scores on the restricted scale. For example, at grade 4 in 2003, one jurisdiction had a full scale mean score 1.2 points below its restricted scale mean score and another jurisdiction had a full scale mean score 0.5 points above its restricted scale mean score. By comparison, the full scale score means for Puerto Rico ranged from 1.9 to 3.4 points below the restricted scale score means (table 3.2).

The fourth row of table 3.4 shows the percentage of jurisdictions for which the full-scale and restricted-scale averages differ by less than 1 scale point. The high percentages indicate the comparability of the mean scores using the two scales. In addition to comparable mean scores on the restricted and full scales, the correlations between the scale scores at grades 4 and 8 in 2003 and 2005 were high (average .99).

Table 3.4

Indicators of agreement between full and restricted scales for the NAEP mathematics, by year and grade

Indicator	2003		2005	
	Grade 4	Grade 8	Grade 4	Grade 8
Difference between full- and restricted-scale estimate for national sample	#	#	0.3*	0.1*
Maximum difference between mean scores for full and restricted scales across jurisdictions	0.5	0.5	0.9	1.6
Minimum difference between mean scores for full and restricted scales across jurisdictions	-1.2	-1.1	-0.7	-1.0
Percent of jurisdictions with full minus restricted-scale score differences of less than 1 point	98.0	98.0	100.0	94.0

Rounds to zero.
* Difference between full- and restricted-scale estimates is statistically significant ($p < .05$).
NOTE: The standard errors do not include linking error and thus may be underestimates of the true standard errors. The statistical procedures and methodology needed to properly estimate the impact of the linking error have not been developed for NAEP.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Conclusions

A set of analyses was conducted to examine whether the NAEP mathematics assessment measured student achievement in Puerto Rico in the same way that it measured student achievement in the nation. Can the results of the 2003 and 2005 NAEP mathematics assessment in Puerto Rico be reported on the NAEP scale? To answer this question, items that were shown to be problematic in Puerto Rico were removed from the full set of items. The two sets of items, the dropped set of administered NAEP items and the kept set of items (those remaining after deleting problematic items), were compared. Results of the comparison indicated that the two sets of items are similar in terms of the distribution of items by content and mathematical ability at grades 4 and 8 in both 2003 and 2005. The distribution of items by item type at grade 4 (but not at grade 8) in the dropped set differed statistically from the distribution of items by item type in the kept set in 2003 and 2005. Items were calibrated, and scales were linked using standard NAEP procedures thus allowing between-scale comparisons.

In the nation, comparisons of the full and restricted scales indicate: (1) mean scale scores agree to within three-tenths of a scale point, (2) the percentage of students performing at various achievement levels agrees to within 1 percentage point, and (3) the correlation among mean scale scores for jurisdictions that comprise the national sample is .99 at grades 4 and 8 in 2003 and 2005. These results indicate the comparability of the full and restricted scales for the nation and the jurisdictions that comprise the national sample.

For Puerto Rico, the difference between the mean scores on the two scales is 2 to 3 scale points, and the difference between the percentage of students meeting a particular achievement level on the two scales is 1 to 2 percentage points. Puerto Rico's mean score was higher on the restricted scale than on the full scale.

What does this mean for presenting and interpreting results of the NAEP 2003 and 2005 administrations in Puerto Rico? Based on these analyses, it is the conclusion of NCES that the full set of NAEP items can and should be used for reporting the Puerto Rico results for both the 2003 and 2005 assessments. The mean scale score for Puerto Rico might be 2 to 3 points higher than reported, but this difference would not change Puerto Rico's ranking among other participating jurisdictions. Similarly for the achievement-level results, the percentages of students at or above a particular achievement level may differ for the two scales by 1 to 2 percentage points.

It would not be appropriate to use the restricted scale for both the nation and Puerto Rico because fewer objectives of the framework could be measured using the restricted scale. Content coverage and stability of estimates over time are key considerations as Puerto Rico moves toward full integration into the NAEP program.

Chapter 4

Puerto Rico Integration into NAEP: Next Steps

In future NAEP administrations, the goal is to include Puerto Rico as part of the national sample and to report the Puerto Rico results with those of other participating jurisdictions. The 2007 administration in Puerto Rico followed procedures used in the 2005 administration. Plans are in place to incorporate Puerto Rico into the NAEP sampling, data collection, and reporting for the 2009 administration. Timely access to NAEP results can assist parents, educators, and policymakers in their efforts to improve student achievement.

2007 Administration

The 2007 NAEP mathematics assessment was conducted January through March. Sampling and data collection procedures were consistent with those of other jurisdictions participating in the 2007 NAEP with two exceptions. First, the entire Puerto Rico administration was in Spanish. Second, students were provided an additional 10 minutes to complete each section of the assessment. These same administration procedures were used in Puerto Rico in 2005.

In addition to the 2007 administration, NAEP conducted a pilot test of new items that, for the first time, included a sample of Puerto Rico students. The results of the pilot test will be used, along with other information, to make final selections of items for the 2009 assessment.

Next Steps

As preparations begin for the 2009 NAEP mathematics assessment, efforts are being made to increase the involvement of educators from Puerto Rico in the development and translation of the mathematics items. These are important next steps toward the goal of fully integrating Puerto Rico into the NAEP program.

• ***Increase the involvement of Puerto Rico representatives in assessment development.*** In preparation for an operational assessment, the NAEP development committee meets to review items before pilot testing, provide advice on the development of scoring guides, and examine pilot test results. The NAEP development committee will include a representative from Puerto Rico who is familiar with the linguistic, cultural, and curricular aspects of mathematics instruction that may be particular to Puerto Rico. Although the NAEP framework remains the basis upon which assessment items are developed, the representative from Puerto Rico could point out items with unfamiliar contexts or language usage that could interfere with accurately measuring students' knowledge and skills.

Second, the NAEP program will include representatives from the Puerto Rico Department of Education in the state item reviews that occur before each pilot test. Since the first NAEP state assessment in 1990, these state item reviews have become an important check on potential fairness issues related to state and regional variations. The comments and concerns of the representatives during these item reviews may result in revisions to items or the exclusion of items from pilot tests.

• ***Include more Puerto Rico educators in the translation process.*** Translation review and verification procedures for the 2005 NAEP administration included two teachers in Puerto Rico, a fourth-grade teacher and an eighth-grade mathematics teacher. Because a Spanish version of the NAEP mathematics assessment is developed for use only in Puerto Rico, some of the experts involved in translation and verification procedures will be natives of Puerto Rico and currently involved in the mathematics education of students in the commonwealth. Continuing to improve the understanding of general issues related to translation and adaptation from one language to another while maintaining the original construct to be assessed is highly desirable.

Plans are in place to incorporate Puerto Rico into the NAEP sampling, data collection, and reporting for the 2009 administration. Increasing Puerto Rico's involvement in the development and review of the assessment at various stages will help in developing an assessment that can be adapted for use in Puerto Rico without altering the construct being measured.

Sampling and Implementation

NAEP sampling procedures in Puerto Rico

The schools and students participating in the NAEP assessment are chosen to be nationally representative. Sampling was conducted in two stages. In the first stage, schools were selected from stratified frames within each jurisdiction. In the second stage, students were selected from within schools. Sampling procedures in Puerto Rico did not differ from the procedures in other jurisdictions because the intent is to include Puerto Rico as part of the national sample in future NAEP administrations.

For the trial NAEP administrations in 2003 and 2005, approximately 100 schools and 3,000 students per grade were sampled in public schools in Puerto Rico. Private schools did not participate in the trial NAEP administrations in Puerto Rico. Table 5.1 presents the sample sizes and target populations for the 2003 and 2005 fourth- and eighth-grade mathematics assessment for public school students in Puerto Rico and the nation. Information is only presented for public school students in the nation, although private school students did participate.

Table 5.1

Student sample size and target populations for NAEP mathematics assessment for public school students in Puerto Rico and the nation, by grade and year

	Puerto Rico		Nation	
Year and grade	Student sample size	Target population	Student sample size	Target population
2003				
Grade 4	3,000	48,000	191,400	3,603,000
Grade 8	2,800	45,000	153,500	3,575,000
2005				
Grade 4	2,800	42,000	168,900	3,745,000
Grade 8	2,800	40,000	159,200	3,662,000

NOTE: Student sample sizes are rounded to the nearest hundred, and target populations are rounded to the nearest thousand.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

School and student participation rates

To reduce the possibility of biased estimates, NCES and the Governing Board established participation rate standards that all jurisdictions, including Puerto Rico, are required to meet. NCES requires a nonresponse bias analysis if the participation rate is less than 85 percent. The Governing Board requires a 70 percent response rate for reporting purposes. In the 2003 and 2005 Puerto Rico assessments, both participation rate standards were met at both grades 4 and 8. Table 5.2 provides school participation rates and Table 5.3 provides student participation rates.

Table 5.2

Number and percentage of public schools participating in NAEP mathematics assessment in Puerto Rico and the nation, by grade and year

	Puerto Rico		Nation	
Year and grade	Number of schools	Weighted percent	Number of schools	Weighted percent
2003				
Grade 4	110	100	6,910	99.8
Grade 8	100	100	5,530	99.6
2005				
Grade 4	110	100	8,700	99.6
Grade 8	110	100	6,460	99.5

NOTE: The numbers of schools are rounded to the nearest tenth.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Table 5.3

Number and percentage of public school students participating in NAEP mathematics assessment in Puerto Rico and the nation, by grade and year

	Puerto Rico		Nation	
Year and grade	Number of students	Weighted percent	Number of students	Weighted percent
2003				
Grade 4	3,000	94	184,300	94
Grade 8	2,800	92	147,600	91
2003				
Grade 4	2,800	95	163,000	94
Grade 8	2,800	93	152,800	91

NOTE: The numbers of students are rounded to the nearest hundred.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

References

Allen, N.J., Donoghue, J.R., and Schoeps, T.L. (2001). *The NAEP 1998 Technical Report* (NCES 2001-509). U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: National Center for Education Statistics.

Allen, N., Jenkins, F., Kulick, E., and Zelenak, C.A. (1997). *Technical report of the NAEP 1996 State Assessment Program in Mathematics* (NCES-1997-951). U. S. Department of Education, Office of Education Research and Improvement. Washington, DC: National Center for Education Statistics.

Embretson, S. and Reise, S.P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R.K., Merenda, P.F., and Spielberger, C.D. (Eds.) (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Holland, P.W. and Wainer, H. (Eds.) (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

No Child Left Behind Act of 2001, Pub. L. 107-110, 115 Stat. 1425 (2002).

The Elementary and Secondary Education Act (ESEA) (Pub. L. 89-10. 79Stat. 77, 20 U.S.C. ch.70)

Appendix

Table A-1

Percentage distribution of dropped, kept, and full sets of NAEP mathematics items at grade 4, by year and item characteristics

Item Characteristic	Dropped	Kept	Full
2003[1]			
Content area			
Numbers sense, properties, and operations	32	47	42
Measurement	19	17	18
Geometry and spatial sense	18	14	15
Data analysis, statistics, and probability	11	11	11
Algebra and functions	21	11	15
Item type*			
Multiple choice	79	57	64
Short constructed response	18	39	32
Extended constructed response	4	5	4
Mathematical ability			
Conceptual understanding	46	37	40
Problem solving	40	34	36
Procedural knowledge	14	29	24
2005[1]			
Content area			
Number properties and operations	43	41	42
Measurement	24	19	20
Geometry	5	17	14
Data analysis and probability	11	11	11
Algebra	16	12	13
Item type*			
Multiple choice	92	56	64
Short constructed response	3	40	32
Extended constructed response	5	4	4
Mathematical complexity			
Low	65	70	69
Moderate	32	30	30
High	3	0	1

* Distribution of dropped items is significantly different ($p < .05$) from distribution of kept items.
[1] The NAEP mathematics framework used in 2005 differed from the framework used in 2003. The names of some of the content areas changed. Mathematical ability was assessed in 2003. Mathematical complexity was assessed in 2005.
NOTE: Details may not sum to total due to rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Table A-2

Percentage distribution of dropped, kept, and full sets of NAEP mathematics items at grade 8, by year and item characteristics

Item Characteristic	Dropped	Kept	Full
2003[1]			
Content area			
Numbers sense, properties, and operations	24	27	26
Measurement	16	15	15
Geometry and spatial sense	18	19	18
Data analysis, statistics, and probability	11	17	15
Algebra and functions	31	22	25
Item type			
Multiple choice	73	63	66
Short constructed response	26	31	30
Extended constructed response	2	6	5
Mathematical ability			
Conceptual understanding	45	34	37
Problem solving	31	35	34
Procedural knowledge	24	31	29
2005[1]			
Content area			
Number properties and operations	23	28	26
Measurement	12	17	16
Geometry	21	21	21
Data analysis and probability	16	13	14
Algebra	28	21	23
Item type			
Multiple choice	75	65	69
Short constructed response	19	31	28
Extended constructed response	5	3	4
Mathematical complexity			
Low	61	64	63
Moderate	39	33	35
High	0	2	2

[1] The NAEP mathematics framework used in 2005 differed from the framework used in 2003. The names of some of the content areas changed. Mathematical ability was assessed in 2003. Mathematical complexity was assessed in 2005.
NOTE: Details may not sum to total due to rounding.
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

Table A-3

Minimum and maximum mean item score for the full and reduced set of NAEP mathematics items by year, grade, and jurisdiction

Jurisdiction	Full set of items		Reduced set of items	
	Minimum mean item score	Maximum mean item score	Minimum mean item score	Maximum mean item score
2003				
Grade 4				
Puerto Rico	.00	.79	.01	.75
Nation	.13	.95	.18	.95
Grade 8				
Puerto Rico	.00	.89	.00	.89
Nation	.12	.94	.14	.94
2005				
Grade 4				
Puerto Rico	.02	.79	.02	.79
Nation	.13	.96	.13	.96
Grade 8				
Puerto Rico	.01	.90	.02	.90
Nation	.07	.94	.11	.94

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2005 Mathematics Assessments.

U.S. DEPARTMENT OF EDUCATION

The National Assessment of Educational Progress (NAEP) is a congressionally mandated project sponsored by the U.S. Department of Education. The National Center for Education Statistics, a department with the Institute of Education Sciences, administers the NAEP. The Commissioner of Education Statistics is responsible by law for carrying out the NAEP project.

Margaret Spellings
Secretary
U.S. Department of Education

Grover J. Whitehurst
Director
Institute of Education Sciences

Mark Schneider
Commissioner
National Center for Education Statistics

Peggy Carr
Associate Commissioner
National Center for Education Statistics

THE NATION'S REPORT CARD

Technical Report of the NAEP Mathematics Assessment in Puerto Rico: Focus on Statistical Issues

September 2007

SUGGESTED CITATION

Baxter, G.P., Ahmed, S., Sikali, E., Waits, T., Sloan, M., and Salvucci, S. (2007). *Technical Report of the NAEP Mathematics Assessment in Puerto Rico: Focus on Statistical Issues* (NCES 2007-462). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, D.C.

CONTENT CONTACT

Emmanuel Sikali • 202-502-7419 • emmanuel.sikali@ed.gov

MORE INFORMATION

- The report release site is http://nationsreportcard.gov
- The NCES web electronic catalog is http://nces.ed.gov/pubsearch
- For ordering information, write to

  U.S. Department of Education
  ED Pubs
  P.O. Box 1398
  Jessup, MD 20794-1398

- Or call toll free 1-877-4ED-Pubs
- Or order online at http://www.edpubs.org

"OUR MISSION IS TO ENSURE EQUAL ACCESS TO EDUCATION AND TO PROMOTE EDUCATIONAL EXCELLENCE THROUGHOUT THE NATION."

www.ed.gov