A Validity Study of the NAEP Full Population Estimates

Larry V. Hedges
Northwestern University

Victor Bandeira de Mello
American Institutes for Research

August 2013
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

This report was prepared for the National Center for Education Statistics under Contract No. ED-04-CO-0025/0012 with the American Institutes for Research. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

2047_08/13

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Phone: 650/ 843-8100
Fax: 650/ 843-8200

Executive Summary

To support an internal evaluation of the impact of changing exclusion rates on reports of statistically significant gains in National Assessment of Educational Progress (NAEP) scores across states, the National Center for Education Statistics (NCES) sponsored research on imputation procedures used to calculate NAEP scores for the excluded students and provided adjusted or full population estimates (FPEs) for the 1996 to 2000 NAEP mathematics gains. The FPE methodology developed by McLaughlin (2005) makes use of information in the student-level NAEP data file, which includes data for students with disabilities (SDs) and English language learners (ELLs) generated from questionnaire responses completed by school staff. In 2009, the task force on FPEs formed by the National Institute of Statistical Sciences and the NAEP-Education Statistics Services Institute (NAEP-ESSI) found that methods used to calculate FPEs were sufficiently sound that there was no identified need for drastic modifications. The task force also recommended that NCES support studies to extend and further validate the methodology for imputing plausible values. The occasion of two special inclusion studies conducted in conjunction with the 2011 NAEP Mathematics Assessment presented just such an opportunity for additional validity research.

Both studies focused on the assessment of otherwise-excluded students by offering accommodations that are not allowed in operational NAEP. One study allowed the use of calculators as an accommodation (in states that permit this accommodation on their state assessments). The other provided students with an inclusion booklet made up of Knowledge and Skill Appropriate (KaSA) blocks that were somewhat easier than standard NAEP blocks. In some states, there were students included in both studies (that is, some students included because of the calculator accommodation and other students included because of the inclusion blocks). In other states, only the inclusion block was offered because the states do not allow a calculator accommodation on their state assessments. After school personnel had finalized their exclusion decisions for the operational assessment, they were asked to reconsider whether excluded students could participate using the calculator or KaSA blocks. If they agreed, these students became participants in the special studies. The data from the special studies were scaled with the data from the operational NAEP assessment, and plausible values were created for the participants in the special studies.

Because these 2011 special inclusion studies yielded a sample of students excluded from operational NAEP for whom both NAEP scaled plausible values and FPEs were available, they provided an opportunity to conduct a validity study of the FPEs. The logic was to compare results from an assessment that included the actual scaled scores for some otherwise excluded students (those who could be included with the special accommodations) with results based on the FPEs.

The total number of operationally excluded students in the 2011 NAEP Mathematics Assessment was 5,049 out of a total sample of 169,452 public school

students (about 3.0 percent). Only 1,197 (23.4 percent of the excluded students) participated in the validity study (891 in the special calculator booklet study and 307 in the inclusion booklet study). This was a much smaller sample size than had been expected. Moreover, the special studies sample differed somewhat from the group of excluded students as a whole in ways that are likely to be related to performance on the assessment. In particular, the students in the special studies sample were rated by school personnel as tending to be among the more able of the excluded student group.

Because of the small sample sizes, the differences between the means of the FPEs and of estimates based on scaled plausible values for the otherwise excluded students (overall and for 14 subgroups) resulted in only one significant difference. However, when 95 percent confidence intervals were constructed to examine for possible bias, the resulting intervals ran from 0 to 10 NAEP points, suggesting that the FPEs may tend to overestimate the actual population parameter. This overestimation is not surprising (and indeed was hypothesized to be the case) because the achievement information on which the FPEs are based is only from assessed students.

It is not clear that FPEs have to be unbiased to be useful, however. Unbiased estimation of unobservable assessment scores is probably an impossible goal in any event. *A principled method that leads to smaller bias in estimating a group that is undercovered in a population may be highly desirable.* Excluding a population subgroup because it cannot be assessed is roughly equivalent (for estimating population averages) to imputing the mean of the assessed population. The special studies samples investigated here scored, on the average, at about the 10th percentile of the assessed population. If we interpret the difference between the average FPEs and scaled plausible values from the special studies as bias, then the results presented here suggest that the bias in imputing the mean of the assessed population is approximately 10 times as large as that in using the FPEs.

When one considers the possibility of improving NAEP population estimates by expanding the pool of tested students, the study also offers some insights. First, because of the small numbers of students successfully recruited into the special studies (and the characteristics of these students, who tended to be rated by their schools as among the most able of the excluded students), the studies suggest that offering the calculator block and KaSA booklet accommodations, by themselves, would not have a substantial impact on national parameter estimates. However, results for the FPE estimates on the *entire* excluded population do show nonnegligible impacts on national parameter estimates. This suggests that if accommodations to include more of the currently excluded students could be found, such accommodations could have a nonnegligible impact on national parameter estimates.

Finally, one can question whether the concept of *full* population estimates is sensible. The reason is that the concept of full population estimates presupposes that there is (at least in theory) an assessment score for every student, including those who are currently excluded from the assessment. If there are students

whom we could not conceive of as participating in the assessment under any conditions, then the concept of "the assessment score they would have obtained if they had participated" may not make sense. One might therefore argue that a group that could never be assessed should be excluded from the definition of the population used to draw inferences. By redefining the population, efforts could focus on developing methods to include as many members of the (newly defined) population as possible in operational assessments and on developing methods to impute scores for those excluded.

CONTENTS

Introduction

In early 2001, to support an internal evaluation of the impact of changing exclusion rates on reports of statistically significant gains across states, the National Center for Education Statistics (NCES) sponsored research on imputation procedures of National Assessment of Educational Progress (NAEP) scores for the excluded students and provided adjusted or full population estimates (FPEs) for the 1996 to 2000 NAEP mathematics gains. The same method was subsequently used to produce FPEs for Grades 4 and 8 in reading, writing, mathematics, and science for each year these assessments were administered since 1994 (McLaughlin, 2005).

The FPE methodology developed by McLaughlin (2005) makes use of information in the student-level NAEP data file, which includes data for students with disabilities (SDs) and English language learners (ELLs) generated by questionnaire responses completed by the school. This file also includes teacher responses about the severity of disability, mastery of English, grade level of instruction, and local testing policies for the student (accommodations). Full population estimates are computed by starting with the achievement distribution of included SDs/ELLs in each state and estimating the difference between excluded and included SDs/ELLs based on information that is available on both sets of students. The method then generates estimates of the plausible achievement scores for students selected for, but excluded from, NAEP participation. More information about how the FPEs are calculated can be found in Appendix A.

Braun, Zhang, and Vezzu (2006) introduced an alternative approach to address the exclusion problem. Their approach is also an imputation procedure based on the same basic assumptions used in McLaughlin (2005). When both approaches were compared, their performances were found to be equivalent (Wise, Le, Hoffman, & Becker, 2006).

In 2009, the task force on FPEs formed by the National Institute of Statistical Sciences (NISS) and the NAEP- Education Statistics Services Institute (NAEP-ESSI) found that methods used to calculate FPEs were sufficiently sound that there was no identified need for drastic modifications. The task force also recommended that NCES support studies to extend and further validate the methodology for imputing plausible values.

In spite of considerable prior research, the NAEP FPEs have not yet achieved operational status, and further validity research has been recommended. The occasion of two special inclusion studies in the 2011 NAEP assessment—designed to determine if students who are excluded from the operational assessment can meaningfully participate if offered one of two special booklets—presented just such an opportunity for additional validity research on the FPEs. The Educational Testing Service (ETS) analyzed the data from these otherwise excluded students, but the analyses were not carried out for the purposes of validating the FPEs.

The purpose of this document is to report a study that capitalized on these inclusion studies to obtain further validity evidence about the FPEs and to better inform a possible decision to move toward operational status for the FPE methodology.

The 2011 Special Inclusion Studies

Fieldwork for two closely related inclusion studies was carried out as part of the 2011 NAEP Mathematics Assessment. The two studies had slightly different student populations and examined different accommodations. One study focused on the use of calculators as an accommodation (in states that permit this accommodation on their state assessments). The other study focused on an inclusion booklet composed of Knowledge and Skill Appropriate (KaSA) blocks.[1] In some states, there were students included in both studies (i.e., some students included because of the calculator accommodation and other students included because of the inclusion blocks). In other states, only the inclusion block was offered because the states did not allow a calculator accommodation on their state assessments. Each study is outlined briefly below, and Figure 1 provides a graphic depiction of the procedure for including students in each of the studies.

The Special Calculator Study

This study focused on students who had an accommodation to use calculators on their state assessment.[2] Therefore, the sample for this study was necessarily drawn from only those states that allowed this accommodation on their state assessments in 2011. (Twenty states allowed the accommodation at Grade 4, and 19 states allowed it at Grade 8.) The procedures were as follows:

[1] Items in the KaSA blocks were intended to reduce item difficulty without compromising content or construct validity. To this end, items were developed for a subset of the NAEP Mathematics Framework objectives that were identified as being appropriate for measuring performance at the lower end of the NAEP scale. Item development for the study was done using the following criteria: (1) Items were required to measure existing NAEP framework objectives. (2) The total item distribution across each of the mathematics content subscales (Number Properties and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra) at each grade should mirror the item distribution called for in the NAEP framework. (3) Objectives from reasoning subtopics would not be used. (4) Multiple choice and short constructed-response item formats would be used. Extended-constructed response items would not be developed. (5) Low- and moderate-complexity items would be developed. No high-complexity items would be developed at either Grade 4 or Grade 8. (6) Item development guidelines from the NAEP Validity Panel's Accessible Block Study would be used to guide item development along with results of cognitive labs. Initial results (from special studies conducted prior to the work reported in this paper) showed that, on average, the *p* values for the KaSA items were higher than for operational NAEP and nonresponse rates were lower.

[2] About 24 percent of all excluded students were excluded because they had an accommodation to use a calculator on their state assessment.

1. NAEP field staff members went through the inclusion decision tree with school staff members using normal operational procedures. If the school staff members indicated that they wanted to exclude a student because the student received or would receive the calculator accommodation on the state test but could not receive that accommodation on NAEP, staff members first determined whether the booklet to which the student had already been randomly assigned contained any calculator blocks. About 38 percent of booklets at Grade 4 and 52 percent of booklets at Grade 8 included at least one block that allows calculators.

    a. If the student was assigned to one of these booklets, staff members pointed this out to the school and asked whether the school would reconsider inclusion given these circumstances. If the school said "yes," the student was counted as included.[3] If the school said "no," the student was included in step 2.

    b. If the student was ***not*** randomly assigned to a booklet with at least one calculator block, then the student was counted as excluded and set aside.

2. After all inclusion decisions had been made, the field staff members informed the school that NAEP was doing special inclusion studies and went through the list of excluded students a second time.

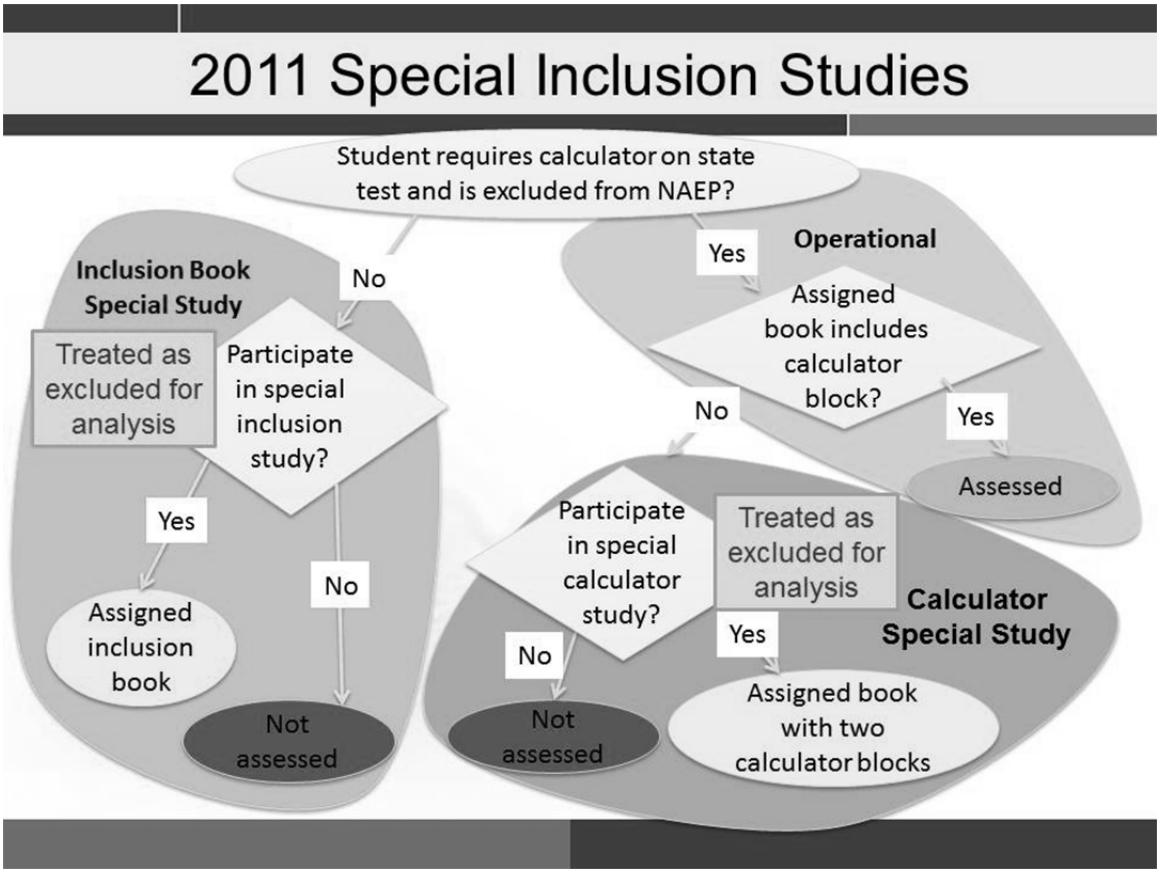
    a. Students who were excluded because they use a calculator accommodation on their state assessment and who were not randomly assigned to a booklet containing at least one calculator block in NAEP were brought up for reconsideration. NAEP offered to give the student a booklet with two calculator blocks if the school would allow the student to participate.

    b. Students who were converted to participating during Step 2 were considered to be in the special study only, and their data were not used operationally. The decision not to use these data operationally seems to have been made out of concern that there are only one or two all-calculator booklets at each grade level, and the sample might become dangerously imbalanced if too many accommodated students accumulated with those few booklets. However, data for these students were processed in the usual way, so their plausible values were available for research purposes.

    c. Because the SD or ELL questionnaires are filled out for all SD and ELL students, whether included or excluded, the information necessary to develop the FPEs for the students in this study was available.

[3] This procedure also was followed in 2009 and appeared to have raised inclusion substantially.

Figure 1. Decision Tree for Inclusion in the 2011 NAEP Special Inclusion Studies



Source: ETS, 2010 Design Summit Briefing Materials.

The Special Inclusion Booklet (KaSA) Study

The second study involved students who were excluded by their schools for ***other reasons than the fact that they receive the calculator accommodation on their state tests***. Therefore, the sample for this study included all the excluded students in states that ***did not*** allow a calculator accommodation on their state assessments in 2011, plus the students in the states that ***did*** allow the calculator accommodation but who were excluded for some other reason.

1. As in the first study, all students were put through the inclusion decision tree first, and the excluded students were set aside.
2. The school was told about the inclusion study, and the "excluded" students were brought up for reconsideration.
3. NAEP staff members offered to test the excluded student with an inclusion booklet composed of two KaSA blocks. If the school then let them participate, the students were assessed with the inclusion booklet.[4]

[4] The KaSA blocks were placed on the NAEP scale in 2011 using a random sample in which some students received booklets with one KaSA block and one operational block.

a. As in the first study, the students in this study were counted as excluded, but their data were processed in the usual way, so plausible values were available for research purposes.

b. Because the SD or ELL questionnaires are filled out for all SD and ELL students, whether included or excluded, the information necessary to develop the FPEs for the students in this study was available.

Validity Studies Based on 2011 Special Inclusion Studies

These 2011 special inclusion studies yielded a sample of excluded students for whom both NAEP scaled plausible values and SD or ELL questionnaire data necessary to develop the FPEs were available. This information provided an opportunity to conduct two related validity studies of the FPEs. The logic of each of these studies was to compare results from an assessment that included the *actual* scaled scores (for the included students and the excluded students who could be included with the special accommodations) with results from an assessment that had scores *adjusted* using the FPEs. Both were done using the Grade 8 NAEP mathematics data.

These studies were conducted in parallel, and the analyses were relatively straightforward. Most of the analyses yielded sets of tables in which the columns were the type of estimate and its standard error (e.g., operational NAEP, operational NAEP plus inclusion study sample, and FPE), and the rows were determined by the parameter of interest (e.g., national mean, national standard deviation, subgroup mean, quantile). Therefore, in describing the studies, we focus on a set of questions that were addressed by both studies.

How Large a Population Do the Special Accommodations Affect?

The information value of any validity studies using these special inclusion studies depends on how substantial a fraction of the (operationally) excluded students can be included using these special accommodations. If this fraction is large, the studies have the potential to provide substantial validity information by providing "valid" NAEP scores for essentially all of the (operationally) excluded population. Conversely, if the fraction of the (operationally) excluded students that can be included using these special accommodations is small, the studies have substantially less potential to provide new validity information.

Therefore, a preliminary question is how large a fraction of the (operationally) excluded population was included with these two special accommodations. NAEP staff members estimated that about 40 to 50 percent (2,000 to 2,500 students) of the currently excluded students could be included with these special accommodations. However, the number of (operationally) excluded students that was included with the special accommodations was smaller than expected—only about 1,200. This limited the questions that could be addressed. If the sample size had been sufficient, it also would have been of interest to determine how these percentages of "includable" students varied across states and NAEP reporting groups.

How Do Estimates of National Population Parameters Based on FPEs Compare With Those Based on the Special Accommodations and Operational NAEP?

The object of the FPEs is to provide estimates of the population parameters that would be obtained if it were possible to include every student in NAEP. Previous evidence has suggested that the FPEs are closer to that ideal than the estimates from operational NAEP (McLaughlin, 2005). These validity studies make it possible to compare both the FPEs and operational NAEP with the parameters of the NAEP score distribution in a population (the operational NAEP population plus those included in the special inclusion sample) that is closer to the ideal of having every student assessed than is true for operational NAEP.

Because the population used for the validity study did not assess all students (some were still excluded), it is particularly useful to compare parameter estimates derived from the FPEs for the subset of the national population that was actually assessed in the validity study. This provides the most direct comparison between FPEs and estimates based on test items.

How Do Estimates of the Population of Excluded Students Based on FPEs Compare With Those Based on the Special Accommodations?

In this study, we focus on the population that was added by the special accommodations and compare the NAEP scale-score distributions estimated from it with those estimated from the FPEs for this same population. The most obvious parameters to consider are the means and standard deviations.

Within the group of operationally excluded students (but who were included in the validity study), the most obvious populations to consider are the national population, the reporting subgroups (e.g., by race/ethnicity), and the states. It is particularly important to examine population subgroups because the fraction of subgroups that are excluded varies substantially. However, the sample size precluded examination of the impact by state.

Note that these comparisons are important, but need to be interpreted with due attention to their implications for a particular validity question. For example, when the fraction of excluded students, the subgroup, or both, are small, differences may not lead to changes in inferences at the national level.

Do the Operationally Excluded Differ From Those Still Excluded Under Special Accommodations?

An additional question is whether the students excluded in operational NAEP differ from those who remain excluded under the special accommodations. To examine this question, we compare the two groups of individuals on items reported on the SD or ELL questionnaires. If the sample of students in the special studies is

representative of excluded students as a whole, then it is plausible that the results of this validity study could provide information about the validity of FPEs for the entire group of excluded students. However, if the sample of students in the special studies differs substantially from the excluded students as a whole, it would be unwise to attempt to generalize the findings to the entire group of excluded students.

Results

The results described in this section are based on included students (N = 164,403), included students plus the special study sample (N = 165,600), and the full population estimates (N = 169,452). The total number of operationally excluded students in the 2011 NAEP Mathematics Assessment was 5,049 out of the total sample of 169,452 public school students (about 3.0 percent). We organize the findings in sections corresponding to the research questions stated in the previous section.

How Much of the Excluded Population Can Use the Special Accommodations?

As noted earlier, the sample size obtained in the special studies was somewhat smaller than was anticipated. We had anticipated that 40 percent to 50 percent of the excluded students (roughly 2,000 to 2,500 students) would have been deemed able to participate in the two special studies. Instead, only 1,197 participated: 891 in the special calculator booklet study and 307 in the inclusion booklet (KaSA) study. Thus, the special study sample consists of only 23.4 percent of the excluded students. This suggests that, although modifications to the assessment, such as the calculator booklet and the special inclusion booklet, could indeed include more students, substantial fractions of students (more than 75 percent of those currently excluded from the assessment) would still be excluded, even if these modifications became operational. Thus, these special accommodations do not, in themselves, resolve the validity problems posed by exclusion of students from the assessment. Moreover, the more fine-grained comparisons we hoped to accomplish (e.g., by state or by crossed reporting categories) were not feasible because of the small sample sizes.

Furthermore, the special studies sample differed somewhat from the group of excluded students as a whole in ways that are likely to be related to performance on the assessment. Because this could compromise the validity of this validity study, we provide information later about the composition of the special study samples in relation to the overall group of excluded students, using information obtained from the SD and ELL questionnaires.

Table 1 shows that the excluded population and the study samples both consist primarily of SDs rather than ELLs. However, the composition of the samples in the special studies is not identical to that of the entire population of excluded students. Table 2 shows that there are somewhat more SDs and somewhat fewer ELLs in the study samples than in the overall population of excluded students. The table also shows that the students in the special studies samples are less likely to be white or Asian and are more likely to be black than are the entire population of excluded students.

Table 1. Disability Status of National Population, Excluded Population, and Special Study Populations for the 2011 Grade 8 NAEP Mathematics Assessment

	Included in Operational NAEP		Excluded From Operational NAEP							
			Total Excluded		Calculator Study		KaSA Study		Not in Either Study	
Subpopulation	Number	Weighted %	Number	Weighted %	Number	Weighted %	Number	Weighted %	Number	Weighted %
SD	17,493	10.5	4,591	91.1	872	97.4	299	96.4	3,420	89.1
ELL	8,586	5.7	814	15.6	78	8.3	27	13.4	709	17.6
SD or ELL	24,704	15.3	5,049	100.0	891	100.0	306	100.0	3,852	100.0
SD who is not ELL	16,118	9.6	4,235	84.4	813	91.7	279	86.6	3,143	82.4
ELL who is not SD	7,211	4.8	458	8.9	19	2.6	7	3.6	432	10.9
SD who also is ELL	1,375	0.9	356	6.7	59	5.7	20	9.8	277	6.7
OVERALL	164,403		5,049		891		306		3,852	

Table 2. Composition of National Population, Excluded Population, and Special Study Populations for the 2011 Grade 8 NAEP Mathematics Assessment

	Included	Excluded Total	KaSA Study	Calculator Study	Excluded but Not in Either Study	
Disability Status						
	10.53	91.1	96.42	97.41	89.08	Percent
	(0.11)	(0.84)	(1.74)	(1.51)	(1.02)	(SE)
SD	[62]	[21]	[18]	[4]	[19]	[df]
	9.6	84.37	86.59	91.74	82.41	
	(0.11)	(1.26)	(4.72)	(2.34)	(1.25)	
SD only	[62]	[15]	[8]	[10]	[37]	
	0.93	6.73	9.83	5.68	6.67	
	(0.04)	(1.15)	(4.54)	(1.88)	(0.97)	
SD and ELL	[59]	[6]	[6]	[6]	[17]	
ELL Status						
	5.73	15.63	13.41	8.26	17.59	
	(0.18)	(1.26)	(4.72)	(2.34)	(1.25)	
ELL	[45]	[15]	[8]	[10]	[37]	
	4.8	8.9	3.58	2.59	10.92	
	(0.16)	(0.84)	(1.74)	(1.51)	(1.02)	
ELL only	[34]	[21]	[18]	[4]	[19]	
Race/Ethnicity						
	53.65	46.11	40.58	43.33	47.31	
	(0.27)	(1.24)	(5.05)	(2.84)	(1.47)	
White	[62]	[62]	[37]	[45]	[48]	
	15.63	23.06	29.16	25.78	21.82	
	(0.25)	(1.01)	(5.42)	(2.95)	(0.99)	
Black	[48]	[56]	[15]	[48]	[52]	
	22.5	23.59	27.68	23.34	23.25	
	(0.31)	(1.46)	(5.41)	(3.51)	(1.49)	
Hispanic	[52]	[30]	[24]	[24]	[48]	
	5.52	4.05	0.75	0.81	5.14	
	(0.17)	(0.46)	(0.55)	(0.35)	(0.58)	
Asian	[32]	[40]	[10]	[14]	[37]	

Note: SE = standard error; df = degrees of freedom.

Focusing on the larger subgroup of SDs only, Table 3 shows that, as might be expected, students in the special study samples are more likely to be described as having mild disabilities and much less likely to be described as having severe disabilities than the overall population of excluded students. Students in the special studies also are much more likely to be accommodated or have a modified assessment in the state assessment and much less likely to have an altered assessment in the state assessment than the overall population of excluded students. It may be particularly important for interpretation that the special studies samples are much more likely to be at or above grade level in mathematics than the general population of excluded SDs. This suggests that the students in the special studies sample would be expected to perform better on NAEP assessment tasks than would the general population of excluded students.

Table 3. Weighted Percentages of Students With Disabilities (SDs) Who Are Not English Language Learners (ELLs) in Various Categories for the 2011 Grade 8 NAEP Mathematics Assessment

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
Student's SD classification	Section 504	6.38	0.61	0.76	1.23	0.43
	IEP	90.20	94.68	96.36	97.00	93.88
	Other or omitted	3.42	4.71	2.88	1.77	5.69
Degree of student's disability	Mild	50.35	23.47	37.26	46.01	15.98
	Moderate	34.20	34.29	55.31	34.86	31.93
	Severe	6.12	29.67	4.27	5.88	38.71
	Omitted	9.32	12.57	3.15	13.24	13.38
How student is included in state assessment	Accommodated	78.86	43.47	66.12	74.20	32.86
	Altered assessment	0.63	39.35	3.92	0.40	53.51
	Modified assessment	4.46	11.69	28.04	23.27	6.87
	No accommodation	13.80	0.83	0.15	1.09	0.83
	Omitted	2.24	4.66	1.77	1.04	5.93
How this student should be included on NAEP test	Not assessed	0.40	57.53	5.46	1.21	78.14
	Assess with accommodations	80.78	37.50	88.70	95.96	16.40
	Assess w/out accommodations	13.87	0.67	0.00	1.08	0.64
	Omitted	4.95	4.30	5.84	1.75	4.82
Student's disability	Autism	3.78	8.92	2.61	3.07	11.14
	Developmental delay	0.26	1.27	1.98	0.30	1.46
	Emotional disturbance	5.73	6.80	6.54	5.61	7.15
	Hearing impairment/deafness	1.53	1.54	0.24	1.72	1.63
	Mental retardation	2.66	24.32	5.95	2.62	32.06
	Orthopedic impairment	0.68	2.45	0.00	1.15	3.06
	Other health impairment	16.66	12.96	13.21	17.08	11.83
	Specific learning disability	58.63	40.22	68.95	65.79	30.35
	Speech or language impairment	8.96	10.96	5.48	7.15	12.56
	Traumatic brain injury	0.25	0.96	3.48	0.25	0.88
	Visual impairment/blindness	0.72	1.31	0.45	0.39	1.64

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
Grade-level student performs in NAEP subject	At or above grade level	29.83	9.01	15.32	16.66	6.30
	1 year below	26.40	16.66	24.72	30.82	12.02
	2 or more years below	30.81	52.48	50.58	37.71	56.63
	No instruction	0.23	2.70	0.00	0.00	3.71
	Do not know or omitted	12.72	19.15	9.39	14.81	21.34
Accommodations	Aide administers test	2.59	2.72	4.74	4.36	2.07
	Braille	0.05	0.24	0.00	0.78	0.13
	Breaks	13.76	9.72	10.23	19.14	7.15
	Calculator	5.67	37.06	42.12	86.57	23.27
	Cueing	4.99	3.85	3.53	3.05	4.10
	Extended time	68.23	37.29	63.67	56.14	29.47
	Large print	0.55	0.46	0.08	0.42	0.50
	Magnification	0.17	0.17	0.00	0.23	0.17
	Read aloud	50.60	42.15	72.54	73.11	30.66
	Respond orally	2.07	3.47	6.36	4.14	2.99
	Sign language	0.26	0.06	0.00	0.09	0.06
	Small group	65.37	42.30	77.55	66.47	32.12
	Template	3.42	4.34	2.22	8.99	3.32
	Other	4.94	12.89	19.70	11.54	12.53

Note: IEP = individualized education program.

Focusing now on the subgroup of students who are ELLs only, Table 4 shows that students in the special studies are much more likely to be included in the state assessment with no accommodation or with simple English as an accommodation than the overall population of ELLs who were excluded from NAEP. Virtually none of the ELLs who participated in the special studies were excluded from the state assessment, but nearly half of the overall population of ELLs excluded from NAEP did not participate in the state assessment. Again, as with the sample of SDs, the ELL students in the special studies samples are more likely to be reported to be at or above grade level in mathematics than are the general population of excluded students who are ELLs. Finally, Table 5 shows that the same tendency for SD-only and ELL-only students in the special studies to be reported as having higher performance in mathematics holds for the smaller group of students who are both ELLs and have disabilities. This reinforces the conclusion that the students in the special studies sample would be expected to perform better on NAEP assessment tasks than would the general population of excluded students.

Table 4. Weighted Percentages of Students Who Are English Language Learners (ELLs) Who Are Not Students With Disabilities (SDs) in Various Categories for the 2011 Grade 8 NAEP Mathematics Assessment

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
How student is included in state assessment	Accommodations	31.44	36.44	53.36	40.59	35.64
	No accommodations	64.63	7.32	0.00	59.41	4.54
	Simple English	1.51	6.45	46.64	0.00	5.49
	Not taken	0.20	47.77	0.00	0.00	52.12
	Omitted	2.21	2.02	0.00	0.00	2.20
How this student should be included on NAEP test	Assess with accommodations	61.51	6.16	0.00	59.41	3.28
	Assess w/out accommodations	33.17	15.19	100.00	40.59	10.91
	Not assess	0.39	77.28	0.00	0.00	84.31
	Omitted	4.93	1.37	0.00	0.00	1.50
Accommodations	Aide administers test	0.38	0.32	0.00	0.00	0.34
	Break	2.39	2.26	0.00	0.00	2.47
	Cueing	0.51	0.41	0.00	0.00	0.45
	Dictionary	13.97	20.16	44.98	16.37	19.55
	Directions in Spanish	0.83	2.87	0.00	0.00	3.13
	Extended time	26.84	28.87	75.84	35.02	26.95
	Items in Spanish	0.29	0.58	0.00	0.00	0.63
	Read aloud	10.98	17.28	30.27	37.42	15.68
	Small group	14.43	20.03	3.67	28.26	20.10
	Spanish version	2.86	2.89	0.00	0.00	3.15
	Other accommodation	1.83	21.30	50.30	2.43	21.43
Receiving instruction in English	No instruction in English	2.27	3.15	0.00	0.82	3.39
	Less than 1 year	9.38	12.64	11.31	12.33	12.71
	1–2 years	7.86	10.11	22.48	3.36	10.10
	2–3 years	4.43	50.61	57.83	15.79	52.39
	3 or more years	66.64	16.93	3.67	65.33	14.57
	Do not know or omitted	9.32	6.45	4.72	2.37	6.74

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
Grade-level performance in NAEP	At or above	33.59	18.48	46.64	67.67	14.69
	1 year below	26.77	8.62	3.67	5.97	8.94
	2 or more years below	23.42	39.51	0.00	21.86	41.85
	No instruction	0.15	2.67	0.00	0.00	2.92
	Do not know or omitted	16.07	30.72	49.69	4.51	31.61
Listening comprehension proficiency	Advanced	46.20	12.72	22.48	64.40	9.40
	Intermediate	33.53	17.14	14.25	12.97	17.48
	Beginning	11.01	42.86	35.47	12.55	44.85
	No proficiency	1.74	20.53	23.09	7.72	21.19
	Do not know or omitted	7.53	6.76	4.72	2.37	7.08
Reading English proficiency	Advanced	34.47	10.01	33.06	59.78	6.36
	Intermediate	41.20	16.58	3.67	16.82	16.99
	Beginning	14.87	37.32	35.47	15.60	38.64
	No proficiency	1.92	29.60	23.09	5.43	31.22
	Do not know or omitted	7.53	6.49	4.72	2.37	6.79
Speaking English proficiency	Advanced	51.14	14.54	22.48	64.40	11.39
	Intermediate	28.18	13.77	3.67	12.20	14.20
	Beginning	11.28	41.81	46.05	15.60	43.18
	No proficiency	1.80	23.39	23.09	5.43	24.44
	Do not know or omitted	7.61	6.49	4.72	2.37	6.79
Writing English proficiency	Advanced	33.59	4.98	22.48	2.66	4.54
	Intermediate	42.68	21.29	14.25	73.94	18.48
	Beginning	14.10	35.39	35.47	15.60	36.53
	No proficiency	1.90	31.08	23.09	5.43	32.83
	Do not know or omitted	7.73	7.26	4.72	2.37	7.63

Table 5. Weighted Percentages of Students Who Are Both English Language Learners (ELLs) and Students With Disabilities (SDs) in Various Categories for the 2011 Grade 8 NAEP Mathematics Assessment

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
How student is included in state assessment	Accommodations	61.64	45.89	69.49	90.92	33.54
	No accommodations	36.66	13.55	26.31	7.89	12.79
	Simple English	0.21	10.15	3.74	0.00	13.10
	Not taken	0.05	29.30	0.46	0.53	39.22
	Omitted	1.44	1.10	0.00	0.67	1.35
How this student should be included on NAEP test	Assess with accommodations	35.09	10.61	3.40	8.61	12.07
	Assess w/out accommodations	59.97	31.58	69.96	88.54	14.54
	Not assess	0.43	56.37	23.37	0.00	72.50
	Omitted	4.51	1.44	3.27	2.84	0.89
Accommodations (ELL)	Aide administers test	1.62	0.07	0.00	0.49	0.00
	Break	4.87	3.52	1.78	7.68	2.95
	Cueing	1.13	1.56	0.98	3.13	1.34
	Dictionary	11.00	5.22	1.85	10.94	4.59
	Directions in Spanish	1.12	0.63	1.25	0.00	0.66
	Extended time	43.49	14.92	11.13	27.96	12.91
	Items in Spanish	0.41	0.08	0.00	0.00	0.10
	Read aloud	29.76	27.17	48.68	73.54	14.85
	Small group	36.41	17.33	9.77	32.89	15.37
	Spanish version	2.11	0.00	0.00	0.00	0.00
	Other accommodations	1.80	24.85	37.88	15.19	24.85
Accommodations (SD)	Aide administer test	2.15	1.44	0.91	3.51	1.10
	Braille	0.05	0.00	0.00	0.00	0.00
	Breaks	10.97	10.37	5.20	10.29	11.14
	Calculator	3.42	20.10	18.84	66.81	10.99
	Cueing	3.79	1.34	0.98	3.37	0.99

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
Accommodations (SD) (continued)	Extended time	59.86	24.15	5.53	32.56	25.19
	Large print	0.06	0.27	0.00	0.00	0.37
	Magnification	0.00	0.00	0.00	0.00	0.00
	Read aloud	47.60	44.31	53.80	80.21	35.78
	Respond orally	1.02	1.57	0.45	2.35	1.58
	Sign language	0.00	0.00	0.00	0.00	0.00
	Small group	57.74	27.70	9.47	36.02	28.71
	Template	0.95	2.51	0.00	9.12	1.55
	Other	4.60	14.80	5.44	8.77	17.37
Receiving instruction in English	No instruction in English	0.40	0.25	1.25	0.00	0.15
	Less than 1 year	4.02	1.44	0.00	4.20	1.10
	1–2 years	4.21	3.42	0.00	2.39	4.13
	2–3 years	0.91	1.53	1.85	0.00	1.78
	3 or more years	80.69	75.39	95.12	78.33	71.90
	Do not know or omitted	9.62	17.72	1.78	14.58	20.67
Grade-level performance in NAEP	At or above	20.97	16.75	28.36	42.75	10.14
	1 year below	22.96	9.08	0.00	7.90	10.46
	2 or more years below	36.33	43.38	65.70	9.98	47.16
	No instruction	0.12	0.19	0.00	0.00	0.26
	Do not know or omitted	19.62	30.60	5.94	39.35	31.99
Listening comprehension proficiency	Advanced	36.50	30.22	63.03	58.54	19.81
	Intermediate	45.01	20.25	12.70	18.20	21.76
	Beginning	9.56	30.21	23.82	7.37	35.65
	No proficiency	0.86	4.38	0.00	0.68	5.76
	Do not know or omitted	8.07	14.94	0.46	15.21	17.01

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
Reading English proficiency	Advanced	22.24	22.98	47.24	39.63	16.13
	Intermediate	45.56	22.10	24.70	26.08	20.93
	Beginning	21.93	30.18	27.59	8.81	34.78
	No proficiency	2.49	5.47	0.00	0.68	7.22
	Do not know or omitted	7.77	19.26	0.46	24.80	20.93
Speaking English proficiency	Advanced	47.94	36.35	61.74	64.15	27.12
	Intermediate	36.94	17.33	13.98	19.68	17.36
	Beginning	5.19	28.57	23.82	0.28	34.85
	No proficiency	1.79	3.59	0.00	0.68	4.69
	Do not know or omitted	8.14	14.17	0.46	15.21	15.97
Writing English proficiency	Advanced	19.91	21.85	8.46	47.60	18.75
	Intermediate	51.39	22.81	49.49	25.36	18.38
	Beginning	18.93	29.17	41.59	9.73	31.17
	No proficiency	2.03	6.43	0.00	2.11	8.22
	Do not know or omitted	7.75	19.74	0.46	15.21	23.49
Student's SD classification	Section 504	1.81	4.45	0.00	30.09	0.00
	IEP	94.88	94.04	100.00	69.91	97.97
	Other or omitted	3.31	1.51	0.00	0.00	2.03
Degree of student's disability	Mild	46.19	20.43	43.06	40.56	13.13
	Moderate	39.91	32.03	14.83	40.32	32.89
	Severe	6.10	31.61	25.42	2.48	38.31
	Omitted	7.79	15.93	16.70	16.64	15.67
How student is included in state assessment	Accommodated	67.38	42.41	33.23	66.96	38.86
	Altered assessment	0.80	38.08	39.41	0.00	45.46
	Modified assessment	14.71	13.79	25.10	21.19	10.67
	No accommodation	14.91	2.63	2.26	11.85	0.85
	Omitted	2.21	3.09	0.00	0.00	4.16

Variable	Category	Included in Operational NAEP	Excluded From Operational NAEP			
			Excluded Total	KaSA Study	Calculator Study	Not in Either Study
How this student should be included on NAEP test	Not assess	0.37	62.50	22.90	1.23	80.59
	Assess with accommodations	78.26	33.65	74.84	91.44	16.02
	Assess w/out accommodations	17.19	0.81	2.26	1.33	0.49
	Omitted	4.17	3.04	0.00	6.00	2.90
Student's disability	Autism	2.44	5.85	0.91	2.15	7.31
	Developmental delay	0.56	0.81	0.00	0.00	1.09
	Emotional disturbance	1.65	3.38	0.00	0.94	4.36
	Hearing impairment/ deafness	1.55	0.70	0.95	1.69	0.47
	Mental retardation	1.86	29.37	0.00	0.00	39.50
	Orthopedic impairment	0.69	1.82	0.00	1.41	2.17
	Other health impairment	4.58	7.92	43.34	3.19	3.70
	Specific learning disability	75.92	55.79	75.23	84.90	47.16
	Speech or language impairment	14.22	7.87	3.41	4.29	9.23
	Traumatic brain injury	0.03	0.54	0.00	0.00	0.73
	Visual impairment/blindness	0.09	0.68	0.00	0.00	0.91
Grade level student performs in NAEP subject	At or above grade level	15.77	12.39	20.64	48.93	3.92
	1 year below	18.73	11.10	2.93	8.96	12.71
	2 or more years below	52.00	46.50	59.74	24.08	49.04
	No instruction	0.04	5.94	0.00	0.00	7.99
	Do not know or omitted	13.45	24.07	16.70	18.03	26.34

Note: IEP = individualized education program.

Our last set of comparisons is based on region. NAEP divides the country into four reporting regions (Northeast, Southeast, Central, and West). The composition of the special studies sample by region of the country differs somewhat from that of the operational NAEP sample. The greatest imbalances are in the West and Southeast. For example, 35.5 percent of the weighted operational sample is from the West, but 52.2 percent of the weighted special studies population is from the West. Similarly, although 24.3 percent of the weighted sample in operational NAEP is from the Southeast, only 7.3 percent of the special studies sample is from the Southeast. The differences are smaller in the Central and Northeast regions, with 21.0 percent and 19.2 percent of the weighted samples in operational NAEP from the Central and Northeast compared with 13.2 percent and 27.0 percent of the weighted special studies samples. These differences seem to have arisen mainly because the different state policies on calculator use determined which states were eligible for the special calculator study. This suggests that it may be important to examine biases in the FPEs by region to determine if these biases appear to be different in the different regions.

Finally, it is not surprising, based on the findings noted earlier, that the NAEP achievement estimate of the average of the entire group of excluded students based on the FPEs is lower than the estimate of the average of the students in the special studies. The overall (weighted) average of the FPEs for all excluded students is about 10 NAEP scale points lower than that of the special studies students; similar differences are observed across subgroups of excluded students formed by gender, race/ethnicity, and region. This finding is consistent with the idea that the students who participated in the special studies would be expected to perform better than the excluded students who did not participate in the special studies.

How Do Estimates of the Population of Excluded Students Based on FPEs Compare With Those Based on the Special Accommodations?

Table 6 is a direct comparison of the parameter estimates based on the FPEs and the scaled plausible values obtained from the special studies sample. The table shows that the differences between the FPEs and estimates based on the special studies are about the same size as the standard error of the estimate or somewhat smaller (about 4 NAEP scale points). The two estimates on the same individuals are correlated, but these correlations are rather small (less than 0.5 in every case). Only one of the differences between estimates based on scaled plausible values and FPEs (for the Southeast region) is large enough to be statistically significant. Although the magnitude of this difference is relatively large (15 NAEP scale points, which is close to three standard errors), it would not be significant at the .05 simultaneous significance level if a Bonferroni adjustment for the 15 tests in Table 6 was applied. When interpreting the size of these differences, it is useful to recall that the national average in operational NAEP is 282.7 and that the 10th percentile in the national distribution is 235.8. Thus, the average student in the special studies scores at about the 10th percentile nationally, and a difference of 4 NAEP scale points is about one tenth of the difference between the average student in these studies and the average (assessed) student nationally.

Table 6. Parameter Estimates Based on Scaled Plausible Values and FPEs for the Special Studies Sample and Various Subgroups for the 2011 Grade 8 NAEP Mathematics Assessment

	Scaled Plausible Values				Imputed FPEs				Correlation		Difference				
Group	**Mean**	***SE***	***N***	***df***	**Mean**	***SE***	***N***	***df***	**Mean**	***SE***	**Diff**	***SE***		***df***	***p* value**
Both studies	237.6	1.76	1,197	35	241.8	3.27	1,197	62	0.32	0.06	−4.2	3.19	−1.324	90	.189
Inclusion booklet study	231.7	3.35	306	35	235.4	4.84	306	34	0.34	0.08	−3.7	4.87	−0.751	61	.455
Calculator booklet study	240.1	2.06	891	22	244.5	3.70	891	62	0.30	0.07	−4.4	3.66	−1.214	84	.228
Gender															
Male	239.0	1.86	783	31	244.4	4.22	783	60	0.31	0.07	−5.4	4.05	−1.340	80	.184
Female	235.2	2.98	414	30	237.2	3.91	414	35	0.32	0.10	−2.0	4.09	−0.494	63	.623
Race/Ethnicity															
White	244.4	2.20	478	43	249.7	3.87	478	58	0.26	0.09	−5.4	3.93	−1.369	90	.174
Black	228.2	3.62	405	27	233.0	5.01	405	29	0.31	0.11	−4.8	5.20	−0.925	52	.359
Hispanic	237.9	3.87	210	26	242.0	7.38	210	17	0.30	0.19	−4.1	7.22	−0.568	27	.575
Asian/Pacific Islander	253.1	11.00	15	11	246.4	11.23	15	11	0.43	0.27	6.6	11.90	0.557	23	.583
American Indian/Alaska Native	223.0	5.14	66	32	219.1	7.08	66	31	0.22	0.21	3.9	7.76	0.505	57	.615
Unclassified	237.8	6.81	23	9	227.6	6.42	23	11	0.10	0.35	10.2	8.89	1.142	19	.268
Region															
Northeast	242.9	2.64	431	44	243.7	5.35	431	51	0.37	0.07	−0.9	5.02	−0.170	74	.865
Southeast	215.0	4.05	118	30	230.0	4.52	118	38	0.30	0.12	−15.0	5.07	−2.957	68	.004
Central	244.2	3.01	155	29	242.4	3.75	155	38	0.26	0.11	1.8	4.17	0.429	66	.669
West	236.4	3.28	493	24	242.4	4.55	493	43	0.28	0.11	−6.0	4.79	−1.246	67	.217

Note: SE = standard error; df = degrees of freedom.

The first panel of Table 6 (rows 1 through 3) show the comparison between average FPEs and average estimates from the scaled plausible values for the combined study sample (the sample for both special studies), the special inclusion booklet study sample only, and the special calculator booklet study sample only. In the two special studies individually and the two special studies combined, the average FPEs were larger than the values obtained from the scaled plausible values. The second panel of the table (rows 4 and 5) shows the estimates for males and females. Again, the average FPEs were larger than the values obtained from the scaled plausible values. However, all of the differences were roughly the size of the standard error of the difference, so not one of them comes close to being statistically significant at the conventional .05 level of statistical significance.

The third panel of the table (rows 6 through 11) shows that the differences among racial and ethnic groups were sometimes positive and sometimes negative. For white, black, and Hispanic students, the average FPEs were larger than the values obtained from the scaled plausible values, while the opposite is true for Asians and Pacific Islanders, American Indians and Alaska Natives, and unclassified individuals. Among regions of the country, the average FPEs were substantially higher than the values obtained from the scaled plausible values in the Southeast and the West, but much less so in the Northeast. The average FPEs were lower than the values obtained from the scaled plausible values only in the Central region. With one exception, the differences between estimates based on scaled plausible values and FPEs were roughly the same size as the standard error of the difference and thus were not statistically significant. The exception is the difference in the Southeast region of the country, which, as noted earlier, is close to three standard errors and would be significant at the 5 percent level if this test were considered alone. However, it is not significant after applying a Bonferroni adjustment based on the 15 comparisons in Table 6.

The large standard errors of the differences, which are a function of the small sample sizes, make interpretation of the differences somewhat ambiguous. On the one hand, for the kinds of students who can be included with the special accommodations studied, these results do not support the conclusion that the bias in the FPEs is different from zero in the nation or in any of the subgroups studied. On the other hand, if we use the differences and their standard errors to construct 95 percent confidence intervals for the possible bias, these results are consistent with a range of possible biases. For example, for the nation as a whole, the 95 percent confidence interval for the bias ranges from about −10 to about +2 NAEP scale-score points. Thus, we cannot rule out (absolute) biases as large as 10 or as small as 0 NAEP scale-score points based on these data.

How Large a Difference Between FPEs and Scaled Plausible Values Is Important?

Although Table 6 gives estimates of the differences between parameter estimates based on FPEs and on scaled plausible values based on the special studies, the interpretation of these differences is difficult without additional context. Here we assume that the estimates based on scaled plausible values estimate the "correct"

quantities, and therefore the differences represent bias induced by the FPEs. Using this interpretation, the question becomes how serious the bias may be. *One might argue that any bias is undesirable, but real data collections face biases, and the consideration in evaluating procedures in real assessments is one of trade-offs and minimizing bias, not one of obtaining exactly unbiased results.*

One perspective for evaluating bias is to compare the biases in the FPEs with the bias in current practice of the operational assessment. One might argue that the operational assessment excludes individuals who, on average, would score low if they could be assessed and that exclusion is roughly equivalent to imputing the mean of the assessed students for each excluded student. From this perspective, the special studies imply that the average bias resulting from this "imputation" is equivalent to the national average of included students minus the average of the special study students, or 283 − 238 = 45 NAEP scale points. The FPE biases overall and in each subgroup considered are estimated to be considerably smaller than that (overall about 10 percent of that value), so from this perspective, the biases in the FPEs estimated here are small.

Another perspective for interpreting the biases is to consider how large the biases would have to be so that they would have a nonnegligible effect on the results of the assessment if the FPEs were used to provide estimates for all excluded students (not just the smaller sample of students in the special studies). Table 7 gives the size of the bias in the FPEs that would be necessary to generate a bias of 0.5, 1.0, and 1.5 NAEP scale points in the national average and in some reporting subgroups.[5] The table shows that a bias of at least 19 points would be necessary to change the national mean by 0.5 scale points. A bias of 38 points is necessary to shift the overall mean by 1.0 NAEP scale points. The table also shows that a bias of 13 points is necessary to change the mean of the most sensitive subgroup (blacks) by 0.5 scale points, and a bias of 26 points is necessary to shift the black subgroup by 1.0 NAEP scale points. (Recall that the observed overall bias estimate and that for the black subgroup were both less than 5 points.) Larger biases are necessary to shift the overall mean or that of other reporting subgroups. Only the bias estimated for the Southeast region (15 scale points) comes close to being large enough to produce a bias of 0.5 points in the average for that reporting subgroup. Therefore, from this perspective, the biases of the FPEs estimated here appear small.

[5] The bias in the estimate of the entire population (included and excluded students) is computed by writing $X_P = w_I X_I + w_E X_E$, where X_P is the entire population mean; X_I and X_E are the means of the included and excluded populations, respectively; and w_I and w_E are their respective weights. Then, note that a change of d points in the excluded population produces a change of $w_E d$ points in the estimate of the mean of the entire population.

Table 7. Bias in FPEs Necessary to Produce a Change of 0.5, 1.0, or 1.5 Points in Overall Averages of Various Groups

	Overall Bias		
Group	0.5	1.0	1.5
Nation	19	38	57
Gender			
Male	16	31	46
Female	26	52	78
Race/Ethnicity			
White	22	44	66
Black	16	26	39
Hispanic	18	37	55
Region			
Northeast	18	37	55
Southeast	22	44	65
Central	18	35	53
West	19	38	56

A more conservative perspective is that the data provided by the special studies have too much uncertainty to support sharp conclusions. For example, the average difference between FPEs and scaled plausible values is −4.2 with a standard error of 3.2, and thus a 95 percent confidence interval for the difference is −11.0 to 2.2. Although even the largest value of 11.0 is not large enough to cause a bias of 0.5 points overall, it is nearly large enough to do so for one reporting subgroup (blacks) and could possibly produce bias of that magnitude in other subgroups not examined here. Moreover, because the special studies sample is different in composition (and possibly in ways that are not observable that might be correlated with achievement) from the entire subpopulation of excluded students, it is difficult to tell if the biases estimated from the special studies also would apply to the entire excluded student population.

How Do Estimates of National Population Parameters Based on FPEs Compare With Those Based on the Special Accommodations and Operational NAEP?

The object of the FPEs is to provide estimates of the population parameters that would be obtained if it were possible to include every student in NAEP. Previous evidence has suggested that the FPEs are closer to that ideal than the estimates from operational NAEP. These validity studies make it possible to compare both the population estimates from FPEs and those from operational NAEP with the parameters of the NAEP score distribution in a population (the operational NAEP population plus those included in the special studies sample) that is closer to the ideal (every student assessed) than is true for operational NAEP.

Table 8 reports the same parameter estimates from the sample included in operational NAEP (excluding special studies sample) and from operational NAEP plus the special inclusion study sample. Estimates from operational NAEP plus the special inclusion study sample are reported first using the scaled plausible values for

the excluded students, then with the FPEs for the excluded students. The table reveals that the estimates including the special studies samples were virtually identical, regardless of how the estimates were derived from the special studies sample. The table also reveals that there are no differences larger than 0.5 NAEP scale points between the estimates based on operational NAEP and the estimates including the special studies sample.

Table 6 showed that there were differences (albeit differences that were typically small in comparison to their sampling uncertainties) in the estimates produced by the scaled plausible values and the FPEs, so both of these findings in Table 8 are a consequence of the small sample size of the special inclusion studies (less than one quarter of the excluded students, which were, in turn, about 3 percent of the assessed sample—i.e., the special study sample was less than 1 percent of the assessed sample).

Table 8. Parameter Estimates for the Operational NAEP Sample and for the Operational NAEP Sample Plus the Special Studies Sample With Scaled Plausible Values and With FPE-Imputed Values

	Operational NAEP (Scaled With Special Studies)			Operational NAEP and Special Studies (Scaled Plausible Values From Special Studies)			Operational NAEP and Special Studies (With Imputed FPEs From Special Studies)		
Group	**Estimate**	***SE***	***N***	**Estimate**	***SE***	***N***	**Estimate**	***SE***	***N***
Overall	282.7	0.19	164,403	282.4	0.19	165,600	282.5	0.20	165,600
Gender									
Male	283.2	0.26	83,305	282.8	0.27	84,088	282.8	0.28	84,088
Female	282.3	0.24	81,098	282.1	0.24	81,512	282.1	0.24	81,512
Race/Ethnicity									
White	292.6	0.21	90,224	292.3	0.21	90,702	292.3	0.20	90,702
Black	261.8	0.42	29,846	261.4	0.42	30,251	261.5	0.49	30,251
Hispanic	269.5	0.47	29,844	269.3	0.47	30,054	269.3	0.48	30,054
Asian/Pacific Islander	302.2	1.03	8,352	302.2	1.03	8,367	302.1	1.03	8,367
Am Indian/Alaska Native	265.9	1.01	3,288	264.8	1.02	3,354	264.8	1.00	3,354
Unclassified	286.3	1.41	2,849	286.0	1.40	2,872	286.0	1.34	2,872
Region									
Northeast	287.0	0.41	36,115	286.6	0.41	36,546	286.6	0.40	36,546
Southeast	279.0	0.33	40,348	278.8	0.34	40,466	278.8	0.32	40,466
Central	285.7	0.41	37,004	285.5	0.41	37,159	285.5	0.41	37,159
West	281.3	0.50	50,936	280.9	0.50	51,429	280.9	0.50	51,429
Quantile									
10th	235.8	0.35		235.3	0.43		235.3	0.39	
25th	258.8	0.22		258.4	0.24		258.6	0.32	
50th	283.7	0.19		283.4	0.21		283.5	0.22	
75th	307.8	0.23		307.6	0.24		307.5	0.30	
90th	328.6	0.27		328.4	0.33		328.4	0.33	
Achievement Level									
Basic and above	72.2	0.22	164,403	71.9	0.22	165,600	72.0	0.27	165,600
Proficient and above	33.5	0.28	164,403	33.3	0.28	165,600	33.4	0.23	165,600
Advanced	7.9	0.15	164,403	7.8	0.15	165,600	7.8	0.16	165,600

Even though the excluded population is small, it does have the potential to affect national parameter estimates. Table 9 shows the same parameter estimates as in Table 8, but it compares the operational NAEP sample with one including FPEs for all excluded students. In this table, there are many differences of more than 1 NAEP scale point between estimates based on operational NAEP and FPEs. Some of these differences may be large enough to have policy implications, particularly as they might play out in state-by-state comparisons.

Table 9. Estimates From the Operational NAEP Sample and From the Operational NAEP Sample Plus Imputed FPEs for Excluded Students

	Operational NAEP			Operational NAEP and Excluded Students (With Imputed FPEs for Excluded Students)		
Subgroup	**Estimate**	***SE***	***N***	**Estimate**	***SE***	***N***
Overall	282.7	0.20	164,403	281.4	0.20	169,452
Gender						
Male	283.1	0.28	83,305	281.5	0.29	86,547
Female	282.3	0.24	81,098	281.3	0.25	82,905
Race/Ethnicity						
White	292.6	0.20	90,224	291.4	0.20	92,390
Black	261.8	0.48	29,846	260.3	0.49	31,270
Hispanic	269.5	0.48	29,844	268.4	0.50	30,840
Asian/Pacific Islander	302.2	1.03	8,352	300.9	1.03	8,595
American Indian/Alaska Native	266.0	0.98	3,288	263.8	0.96	3,432
Unclassified	286.4	1.34	2,849	284.9	1.41	2,925
Region						
Northeast	287.0	0.39	36,115	285.7	0.40	37,404
Southeast	278.9	0.32	40,348	277.7	0.32	41,352
Central	285.6	0.42	37,004	284.1	0.42	38,220
West	281.3	0.50	50,936	280.0	0.52	52,476
Quantile						
10th	235.8	0.37		233.3	0.39	
25th	258.9	0.33		257.4	0.34	
50th	283.7	0.25		282.7	0.21	
75th	307.7	0.24		307.0	0.27	
90th	328.5	0.33		328.0	0.34	
Achievement Level						
Basic and above	72.3	0.27	164,403	71.0	0.29	169,452
Proficient and above	33.5	0.24	164,403	32.8	0.23	169,452
Advanced	7.8	0.16	164,403	7.6	0.16	169,452

Validity Considerations of the Validity Study

One threat to what might be called the statistical validity of the study is that the sample size may not be large enough to provide adequate statistical power or precision for the estimates compared. To evaluate this, it is important, as a first step, to examine the sample sizes obtained and determine if the estimates for the operationally excluded subgroup are precise enough to draw conclusions. The logical framework of this study is that of an equivalence (not superiority) study. That is, we conclude that FPEs are valid if the estimates based on them do not differ from those based on scaled plausible values for students in the inclusion samples. Consequently, we must set the smallest difference that is meaningful and determine whether the sample size will yield adequate statistical power to detect such differences. As a guideline, we suggest using the convention of 80 percent power at the 5 percent significance level.

A crude precision analysis can be done by computing the standard error of the difference between the estimate of a population parameter (e.g., the mean) based on full population methods Y_{FPE} and the same estimate based on scaled plausible values Y_{SPV}

$$S = \sqrt{S_{FPE}^2 + S_{SPV}^2 - 2S_{FPE}S_{SPV}r},$$

where S^2_{FPE} and S^2_{SPV} are the variances of Y_{FPE} and Y_{SPV}, and r is an estimate of the correlation between them. A crude estimate of the power to detect a true difference between Y_{FPE} and Y_{INC} of size δ is

$$p = 1 - \Phi(c - \delta/S) + \Phi(-c - \delta/S),$$

where c is the appropriate critical value of the standard normal distribution and $\Phi(x)$ is the standard normal cumulative distribution function.

The question of how large the difference δ must be to be meaningful is more difficult. We used the approach of studying how large δ must be to produce a consequential difference in assessment scores, as in Table 7. Using the standard errors from Table 6, we evaluate the power to detect the smallest bias that would lead to a change in national or subgroup means by 0.5 and 1.0 NAEP scale-score points in Table 10. For the nation and all of the subgroups considered, the power to detect a bias large enough to change the average estimate by 1.0 NAEP scale-score points is essentially 1.0; thus, these studies appear adequately powered to detect biases large enough to produce a change of 1.0 NAEP scale-score point.

The situation is somewhat different with respect to biases large enough to produce a change of 0.5 NAEP scale-score points. In the black and Hispanic subgroups, the power to detect such a change is only about 70 percent. Thus, the special studies cannot be considered definitive in ruling out such biases in the black and Hispanic reporting subgroups. Note also that, although the point estimates of bias for these two subgroups were less than 5 NAEP scale-score points, the upper ends of the

95 percent confidence intervals for the bias estimates in these two groups (17.2 and 20.8, respectively) do exceed the threshold for bias that could cause a 0.5 NAEP scale-score change in national estimates for each of those groups. In other words, the power of these validity studies is not high enough to rule out biases that could change national estimates of the mean in the black and Hispanic reporting subgroups by as much as 0.5 NAEP scale-score points.

Table 10. Power to Detect a Bias in FPEs That Could Produce a Change in Overall Averages of 0.5 or 1.0 Points in Various Groups

	To Detect Overall Bias of 0.5		To Detect Overall Bias of 1.0	
Group	**FPE Bias**	**Power**	**FPE Bias**	**Power**
Nation	19	1.00	38	1.00
Gender				
Male	16	0.98	31	1.00
Female	26	1.00	52	1.00
Race/Ethnicity				
White	22	1.00	44	1.00
Black	13	0.71	26	1.00
Hispanic	18	0.70	37	1.00
Region				
Northeast	18	0.95	37	1.00
Southeast	22	0.99	44	1.00
Central	18	0.99	35	1.00
West	19	0.98	38	1.00

Note: These computations assume a two-sided 5 percent nonsimultaneous significance test.

There also are two threats to internal validity that can be characterized as selection threats. If the school personnel making exclusion decisions know that this is part of a special study, then biases might arise because of experimenter demand characteristics (see Orne, 1962) or Hawthorne effects (Mayo, 1949).[6] We believe that the data collection plan did *not* explicitly characterize this as part of a special study, which should minimize that validity threat.

The second selection threat is that the school personnel might be motivated to exclude from the operational assessment the students that they believe will perform most poorly. Because they are told that the initially excluded students will not be part of the operational assessment, they have no incentive to exclude students they believe will perform most poorly from the validity study. However, any tendency to exclude students whom they believe will perform most poorly from the *operational*

[6] Experimenter effects refer to experimental results that are biased as a result of the study participants' desire to please the researcher. Hawthorne effects are similar. In a classic study of worker productivity at the Western Electric Hawthorne factory, it was shown that the results were less due to the interventions that were put in place than the fact that the workers were being studied, which seemed to increase productivity in and of itself.

assessment could mean that the validity study sample may include students who could have been included in the operational assessment, but who were systematically excluded because they were expected to have poorer performance than those included.

The implications for performance of the sample of operationally excluded students who are in the validity study are unclear because these two factors work in opposite directions. Assuming that, in general, students who *could properly* participate in the operational assessment will perform better than those who could not, adding these excluded students to the validity study sample might artifactually elevate the performance of the students in the validity study. However, if school personnel are correct that the operationally excluded students perform more poorly than included students, they may also perform more poorly than the properly excluded students, which would artifactually reduce the performance of the students in the validity study sample.

The basic validity question is whether the excluded students who participated in the special studies differ from other excluded students in unobservable (or at least unmeasured) ways that are correlated with achievement (holding constant the observables used in creating the FPEs). The fact that the results in Table 6 suggest that estimates of average achievement based on scaled plausible values are slightly smaller than those based on FPEs suggests that this may be the case.

More elaborate statistical modeling to estimate the expected performance of the excluded students also would be possible. For example, suppose that the excluded students are modeled to be the lower tail (the lowest x% of the distribution, where x is the exclusion rate) of the achievement distribution. We could use the assumption of a distribution shape (e.g., normal) to obtain the expected average (and even standard deviation) of the excluded group. Such an analysis would not, however, resolve whether the poorer performance of excluded students was a consequence of proper exclusion (which is consistent with excluded students having poorer performance) or improper exclusion (excluding students who could have participated but who were excluded because they were expected to have poorer performance).

Conclusions

The special inclusions studied here are disappointing in that they made it possible to include in the assessment only about a quarter of the excluded students and less than 1 percent of the total sample. Moreover, the students they made it possible to include appear to be among the most able of the excluded students—those who were "almost able" to be included without the special accommodations. The cost of these special accommodations seems relatively large for the potential benefit achieved.

In general, it appears that the FPEs may tend to overestimate the results based on scaled plausible values in the special studies, although these differences are far from statistically significant. This is not surprising (and indeed was hypothesized to be the case) because the achievement information on which the FPEs are based is from assessed students. Presumably, there are reasons that students are not assessed, and

not all of these depend on observable (or at least observed) characteristics. Thus, any assessed student whose observed characteristics are equivalent to a student who is not assessed differs on some characteristics that are not observed. If this is so, and if these unobserved characteristics are correlated with (also unobserved) assessment scores, then the FPEs would be biased estimates of the assessment scores. More specifically, it is plausible that the unobserved information leading to exclusion is negatively related to assessment scores. If so, then FPEs would overestimate the performance of excluded students.

It is not clear that FPEs have to be unbiased to be useful, however. Unbiased estimation of unobservable assessment scores is probably an impossible goal in any event. *A principled method that leads to smaller bias in estimating a group that is undercovered in a population may be highly desirable.* Excluding a population subgroup because it cannot be assessed is roughly equivalent (for estimating population averages) to imputing the mean of the assessed population. The special studies sample investigated here scored, on the average, at about the 10th percentile of the assessed population. If we interpret the difference between the average FPEs and scaled plausible values from the special studies as bias, then the results presented here suggest that the bias in imputing the mean of the assessed population is approximately 10 times as large as that in using the FPEs.

The composition of the special studies sample appears to include more able students than the average of the excluded student population. If this is true, then the difference between the (unobserved) ability of the entire excluded population and the FPEs (the bias in the FPEs) could be larger for the entire excluded population than for the special studies sample. Although the special studies provide no empirical evidence about the size of that bias, it is difficult to imagine that it could be larger than the bias implied by imputing the mean of the assessed population for these values.

These studies suggest that the calculator block and KaSA booklet accommodations, by themselves, will not change the number of included students enough to have a substantial impact on national parameter estimates. However, results for the FPE estimates on the entire excluded population do show nonnegligible impacts on national parameter estimates. This suggests that if accommodations to include more of the currently excluded students could be found, such accommodations could have a nonnegligible impact on national parameter estimates. Moreover, because FPEs appear to overestimate estimates based on scaled plausible values, the impact of including currently excluded students would likely be even larger than that estimated by the FPEs.

It is important to remember that these special studies are relatively small, and consequently their results have considerable sampling uncertainty that makes it difficult to draw sharp conclusions. The sampling uncertainty made it infeasible to carry out many analyses that would have been desirable. For example, it would be useful to see if patterns of bias were reasonably constant across states and across all reporting groups, but it was not meaningful to conduct these analyses. A fair conclusion is that the sampling uncertainty is so large that any conclusions drawn from this study must be done with extreme care.

It may be useful to question whether the concept of *full* population estimates is sensible. The reason is that the concept of full population estimates presupposes that there is (at least in theory) an assessment score for every student, including those who are currently excluded from the assessment. If there are students whom we could not conceive as participating in the assessment under any conditions, then the concept of "the assessment score they would have obtained if they had participated" may not make sense. Moreover, it is impossible that any empirical methods could be developed to impute assessment scores for a group that could never have assessment scores—no empirical information about assessment scores could exist for that group. Consequently, it will never be possible to validate methods of imputing assessment scores for a group that could never be assessed. One might therefore argue that a group that could never be assessed should be excluded from the definition of the population used to draw inferences. By redefining the population, efforts could focus on developing methods to include as many members of the (newly defined) population as possible in operational assessments and developing methods to impute scores for those excluded. Of course, there is a problem in identifying the group that should be defined as not (ever) assessable. Nevertheless, it may be worth attempting to develop at least provisional definitions of such a group.

This suggests a concept of *expanded* population estimates (rather than full population estimates) that corresponds to estimating the assessment scores that could be obtained by all students who could participate in the assessment under conditions of special accommodations. One virtue of this definition is that every student in the inference population could be assessed under *some* accommodations (including accommodations that might be infeasible under operational conditions because of time or cost constraints). Because it would be possible to obtain assessment scores for every student in the population, empirical methods could, in principle, be used to develop imputations for any students in this population who are excluded from the operational assessment (perhaps in special studies involving extensive accommodations). Moreover, it would be possible to empirically validate such methods.

References

Braun, H., Zhang, J., & Vezzu, S. (2006). *Evaluating the effectiveness of a full-population estimation method* (Unpublished paper). Educational Testing Service.

Mayo, E. (1949). *Hawthorne and the Western Electric Company*, *The social problems of an industrial civilisation*. New York: Routledge.

McLaughlin, D. H. (2000). *Protecting state NAEP trends from changes in SD/LEP inclusion rates* (report to the National Institute of Statistical Sciences). Palo Alto, CA: American Institutes for Research.

McLaughlin, D. H. (2005). *Properties of NAEP full population estimates*. Palo Alto, CA: American Institutes for Research.

National Institute of Statistical Sciences. (2009). *Final report: NISS/NESSI Task Force on Full Population Estimates for NAEP*. Research Triangle Park, NC: Author.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.

Wise, L. L., Le, H., Hoffman, R. G., & Becker, D. E. (2006). *Testing NAEP full population estimates for sensitivity to violation of assumptions: Phase II* (Draft Technical Report DTR-06-08). Alexandria, VA: HumRRO.

Appendix A. Procedures for Calculating Full Population Estimates

McLaughlin (2000) introduced a method to estimate the achievement of the subset of the students with disabilities (SDs) and English language learners (ELLs) excluded by NAEP. The method relies on the NAEP SD and ELL questionnaires, descriptive surveys that are filled out by a teacher or knowledgeable staff person for each student with a disability and each English language learner selected to participate in NAEP—whether or not these students actually participate in NAEP or are excluded on the grounds that NAEP testing would be inappropriate for them.

The basic assumption is that excluded students in a given state with a particular profile based on student and school demographic characteristics and information from the SD and ELL questionnaires will, on average, be at the same achievement level as students with disabilities and English language learners in that state who participated in NAEP and had the same profile of demographic characteristics and information on the SD and ELL questionnaires. McLaughlin called this the profile matching method. Since the scores resulting from this procedure provide estimates that now include all of a state's SDs and ELLS, they are called *full population estimates* (FPEs).

No student takes the entire NAEP assessment. Instead, a student takes a random sample of blocks of items drawn from the entire item set for a given assessment. The items each student takes are used to compute five sets of what are called *plausible values*.[7] These are then used to compute estimates of performance for the entire population of students as well as congressionally mandated subgroups of students (e.g., males and females).

In computing the FPEs, plausible values for the composite NAEP scale in each grade and subject are computed first for all excluded ELLs in the NAEP public school sample, and second, separately, for all excluded SDs in the sample who are not also ELLs. Data for students who are neither ELL nor SD are not used in the process. The plausible values are constructed in three steps.

1. **Predictor preparation.** Predictive demographic information and questionnaire responses, which are available for both included and excluded ELLs and SDs, are extracted from the NAEP file and recoded to maximize predictive power. Stepwise regression is used to remove predictors possessing no significant power in predicting plausible values for included ELLs (or SDs) and to remove predictors that are too highly correlated with other predictors.
2. **Estimation of the mean expected score for each excluded student**. A single pooled within-state linear regression is carried out to estimate the coefficient for each of the predictors created in step 1 in predicting the scores of included ELLs

[7] The procedures in this paper used five plausible values, but the estimation procedure has been changed for the 2013 NAEP assessments and now generates 20 plausible values. Future versions of the software for generating FPEs will be updated to reflect this change.

(or SDs).[8] The regression intercept is adjusted separately for each state so that the mean predicted score for included ELLs (or SDs) matches their observed mean in each state. The resulting coefficients are used to impute an estimate for each excluded ELL (or SD).

3. **Estimation of imputation error variance and generation of five random plausible values for each excluded student**. Five plausible values are generated for each excluded student by adding to the estimate obtained in step 2 random normal deviates with three components of variance: (1) average variation among the five NAEP plausible values for included ELLs (or SDs), (2) average regression error due to the imperfect linear regression prediction in step 2, and (3) sampling error introduced in matching the included ELL (or SD) mean in the state.

One of the difficulties that the FPE procedure has had to deal with is that the set of questions that comprise the NAEP SD and ELL questionnaires have changed from year to year. As a result, the prediction equations change from NAEP administration to administration. While this fact does not diminish the utility of the FPE procedure, it does mean that the fit of regression results to the data can vary over time. Table A-1 below lists separately the variables used in the NAEP 2011 Grade 8 reading and mathematics FPE regressions for ELLs and SDs.

[8] A student's "score" is defined as the mean of the five plausible values for that student.

Table A-1. Variables Used in the Linear Regressions for Grade 8 Reading and Mathematics: 2011

		Mathematics Grade 8		Reading Grade 8	
Variable	**Description**	**SD**	**ELL**	**SD**	**ELL**
	Items from the ELL questionnaire				
XL04501	What is this student's ELL classification?				•
XL03801	How is student included in state assessment?		•		•
XL03901	Extended time (allowed for all subjects)				
XL03902	Small group (allowed for all subjects)				
XL03908	Test items read aloud in English				
XL03905	Breaks during testing (allowed for all subjects)				
XL03909	Must have an aide administer test		•		
XL03910	Cueing to stay on task				
XL03906	Bilingual dictionary w/out definitions in any language				
XL03911	Read directions aloud in Spanish				•
XL03912	Test items read aloud in Spanish (math & science)				•
XL03913	Spanish/English version of the test (math & science)		•		
XL03914	Student receives other accommodations				
XL04001	How should this student be included on NAEP test?				
XL04002	If student ineligible for NAEP, record admin. code				•
XL04101	How long has student been receiving instruction in English?				•
XL04201	Grade level of performance in NAEP subject		•		•
XL04301	Student's English proficiency: listening comprehension in English		•		•
XL04302	Student's English proficiency: Speaking English		•		
XL04303	Student's English proficiency: Reading English		•		•
XL04304	Student's English proficiency: Writing English		•		•
	Items from the SD questionnaire				
XS04701	Why is this student classified as SD?	•		•	
XS04801	How is student included in state assessment?	•	•	•	•
XS04901	Extended time (allowed for all subjects)		•	•	•
XS04902	Small group (allowed for all subjects)	•		•	
XS04907	Test items read aloud in English	•		•	•
XS04905	Breaks during testing (allowed for all subjects)				
XS04908	Must have an aide administer test		•		
XS04909	Responds orally to a scribe	•			
XS04910	Large-print version of the test	•			
XS04911	Magnification equipment	•			
XS04912	Uses a calculator for all sections (math only)	•			
XS04913	Uses template/special equip./preferential seating	•		•	

		Mathematics Grade 8		Reading Grade 8	
Variable	**Description**	**SD**	**ELL**	**SD**	**ELL**
XS04914	Cueing to stay on task				
XS04915	Presentation or response in braille				
XS04916	Presentation or response in sign language	•			
XS04917	Student receives other accommodations	•			
XS05001	How should this student be included on NAEP test?	•			
XS05002	If student ineligible for NAEP, record admin. code				
XS05101	Student's identified disability: Specific learning	•		•	•
XS05102	Student's identified disability: Hearing impairment	•			
XS05103	Student's identified disability: Visual impairment			•	
XS05105	Student's identified disability: Mental retardation	•	•	•	
XS05106	Student's identified disability: Emotional disturbance	•			
XS05107	Student's identified disability: Orthopedic impairment		•	•	
XS05108	Student's identified disability: Brain injury	•		•	
XS05109	Student's identified disability: Autism	•	•		
XS05110	Student's identified disability: Developmental delay				
XS05111	Student's identified disability: Other health	•	•		
XS05104	Student's identified disability: Speech impairment	•		•	•
XS05112	Student's identified disability: Other-write-in				
XS05201	Degree of student's disability	•	•	•	
XS05301	Grade level student performs in the NAEP subject	•	•	•	•
	Student and school characteristics				
IEP	Student classified as having a disability		•		•
DMIN	Student is not white	•		•	•
DSEX	Student gender	•	•	•	•
SLUNCH	National School Lunch Program eligibility	•	•	•	•
PCTBLK	School-level percentage of black students	•	•		•
PCTIND	School-level percentage of American Indian students	•	•		•
PCTHSP	School-level percentage of Hispanic students	•	•	•	
READVAR	School-level state test scores—Reading	•		•	•
MATHVAR	School-level state test scores—Math	•	•	•	
SENROL8	School enrollment		•		

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2011 Reading and Mathematics Assessments.