Examining NAEP Achievement in Relation to School Testing Conditions in the 2010 Assessments

Ina V.S. Mullis
Boston College

George W. Bohrnstedt
American Institutes for Research

Anna Corinna Preuschoff
Boston College

Illiana de los Reyes
American Institutes for Research

Fran Stancavage
American Institutes for Research

Michael O. Martin
Boston College

February 2012
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.

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:

Albert E. Beaton
Boston College

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
University of California, Berkeley

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado at Boulder

Ina V.S. Mullis
Boston College

Scott Norton
Louisiana Department of Education

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado at Boulder

David Thissen
University of North Carolina, Chapel Hill

Karen Wixson
University of North Carolina, Greensboro

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

CONTENTS

Overview of Validity Concern

The purpose of this study was to examine National Assessment of Educational Progress (NAEP) testing conditions in schools and investigate whether being assessed in less than optimal testing conditions is associated with lower student achievement on the assessments. It is well known that NAEP has expended considerable effort to ensure high quality in data collection by developing standardized materials and survey operation procedures and using well-trained professional administrators provided by Westat, the contractor for assessment administration. However, at the higher grades, NAEP can sometimes sample as many as 90 students per school, and schools are allowed to minimize the disruption associated with pulling students out of classrooms by having all of these students assessed at one time, using an auditorium, lunchroom, library, or media lab. This policy is intended to encourage sampled schools to participate, but may have unintended negative consequences for the testing conditions experienced by these students.

For example, it seemed plausible that seating arrangements in nonclassroom settings might hamper students from being able to remain focused on the assessment and might not provide them enough space to work. Also, venues other than classrooms might be susceptible to distractions—for example, a play rehearsal in the auditorium or the clanging sounds of the kitchen staff preparing for lunch.

More broadly, the primary research questions for this study were the following:

1. Are NAEP testing conditions in schools consistent with best assessment practices?
2. If students are assessed in crowded, noisy, or otherwise disruptive conditions, is this associated with lower performance?

As the study progressed, a third question became important:

3. If disruptive testing conditions are related to lower performance, how does this affect NAEP's estimates of the gaps in average scores between advantaged and disadvantaged groups of students?

To collect information for this study, the session debriefing form completed by Westat assessment administrators was expanded for 2010 to include a new set of questions developed by the study authors. The new questions addressed the extent to which the seating arrangements in sessions provided space for students to work and staff to monitor the assessment, the adequacy of lighting and heating/cooling, and the amount of noise, visual distractions, and disruptions to the sessions. The complete 2010 session debriefing form is shown in Appendix A.

In 2010, NAEP assessed civics, U.S. history, and geography at grades 4, 8, and 12. In addition, smaller numbers of students participated in an accessible booklet study in mathematics (grades 4 and 8) and in a writing pilot (grade 4). The civics booklets were given in one set of sessions, sometimes spiraled with the accessible mathematics and pilot writing booklets, while the U.S. history and geography booklets, which have an older design in terms of the placement of the background questions, were spiraled in a separate set of sessions. At each grade, therefore, NAEP ran two sets of sessions covering three operational subjects—the civics sessions and the U.S. history and geography sessions.

As shown in Table 1, results for the testing conditions study are based on 1,316 sessions and 16,698 students at grade 4; 1,039 sessions and 26,372 students at grade 8; and 913 sessions and 25,182 students at grade 12. This includes the students assessed in civics, U.S. history, and geography. The students who participated in the writing pilot or accessible mathematics study were not included. Also, results are based only on regular sessions, not including separate accommodated sessions (that is, sessions for students who had "small group administration" or "individual administration" as an accommodation), and not including makeup sessions.[1]

Table 1. Number of Sessions and Students in Testing Conditions Study of the NAEP 2010 Civics, U.S. History, and Geography Assessments

	Number of Sessions	Number of Students
Grade 4	1,316	16,698
Grade 8	1,039	26,372
Grade 12	913	25,182

Student Achievement Under Various Testing Conditions

Table 2 contains a series of panels shown on two pages that correspond, in general, to the order of questions in the session debriefing form. More specifically, Table 2 provides the percentage of students and their average achievement for the question response categories in the session debriefing form.

[1] Consistent with initial study plans, Westat did not forward the debriefing information for separate accommodated sessions (individual or group). The makeup session data were excluded after analyzing the data and determining that the makeup session sizes were typically very small. There were relatively few makeup sessions, although the numbers increased across the grades: grade 4 had 41 makeup sessions (0.5 percent of the students), grade 8 had 112 (1.5 percent of the students), and grade 12 had 311 (6.5 percent of the students). Average makeup session sizes were three students at grade 4, four students at grade 8, and five students at grade 12.

Table 2. Percentage of Students and Average Achievement (in Normits) in the Testing Conditions Study of the NAEP 2010 Civics, U.S. History, and Geography Assessments by Characteristics of the Testing Session, as Reported on the Session Debriefing Form

		Session Size							
		<20		20–40		41–60		>60	
	Overall Average Achievement	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**
Grade 4	.07	35	.07	62	.06	3	.22	1	~
Grade 8	.08	13	.11	51	.05	35	.11	2	.43
Grade 12	.05	20	.02	44	.01	31	.08	5	.27

	Session Location									
	Classroom		**Auditorium**		**Lunchroom**		**Library**		**Other**	
	Percent Students	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**
Grade 4	85	.07	#	~	4	.13	5	-.07	6	.14
Grade 8	42	.06	4	.11	27	.12	17	.07	11	.09
Grade 12	34	.04	11	.15	19	.03	17	-.03	20	.09

	Session Day									
	Monday		**Tuesday**		**Wednesday**		**Thursday**		**Friday**	
	Percent Students	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**
Grade 4	7	.29	32	.06	27	.02	26	.03	8	.19
Grade 8	7	.13	25	.01	31	.08	28	.09	9	.24
Grade 12	7	.13	29	.02	28	.09	28	-.02	9	.13

	Grade 4		**Grade 8**		**Grade 12**	
Original Debriefing Form Questions	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**
Problems setting up	8	.01	9	.04	8	-.14
Problems getting students there	10	-.06	19	-.07	37	-.03
Problems with timing	2	~	3	.02	3	-.22
Problems with materials	2	~	1	~	1	~
Student refusals	2	~	6	.08	24	.09
Students left session	32	.06	36	.08	38	.02
Problems with NAEP calculators	#	~	1	~	#	~
Problems with accommodations	1	~	1	~	1	~
Students still working	69	.06	56	.03	52	.03
Problems with location	4	.01	9	-.04	10	-.01
Interruptions	10	.07	16	-.01	19	.00
Other	3	.05	2	~	3	.07

Table 2. Percentage of Students and Average Achievement (in Normits) in the Testing Conditions Study of the NAEP 2010 Civics, U.S. History, and Geography Assessments by Characteristics of the Testing Session, as Reported on the Session Debriefing Form (Continued)

	Agree a Lot		Agree a Little		Disagree a Little		Disagree a Lot	
New Study Questions	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**
Adequate Space for Students to Work								
Grade 4	78	.08	15	.05	4	-.03	2	~
Grade 8	79	.10	12	.01	6	.12	2	-.13
Grade 12	83	.07	11	-.02	4	-.09	1	~
Ample Space to Monitor Students								
Grade 4	86	.08	10	.01	3	-.08	1	~
Grade 8	85	.09	9	.06	4	.15	1	~
Grade 12	85	.05	9	.12	4	-.15	2	~
Lighting Adequate								
Grade 4	97	.07	3	.01	#	~	0	~
Grade 8	96	.08	2	.16	1	~	#	~
Grade 12	96	.05	2	.06	1	~	#	~
Temperature Comfortable								
Grade 4	82	.08	12	.05	4	.00	1	~
Grade 8	80	.11	13	.02	5	.00	2	~
Grade 12	80	.08	12	-.06	5	-.14	2	~
Room Noisy Because School Activity								
Grade 4	2	~	3	.14	4	-.29	90	.08
Grade 8	2	-.12	10	.02	9	-.01	77	.11
Grade 12	2	~	9	-.13	7	.06	81	.07
Visual Distractions								
Grade 4	1	~	3	-.08	4	.01	92	.07
Grade 8	2	-.06	5	.03	6	.00	86	.10
Grade 12	1	~	4	-.21	5	.05	89	.06
Numerous School Disruptions								
Grade 4	1	~	2	~	4	-.10	92	.07
Grade 8	2	.00	6	.05	9	.07	82	.09
Grade 12	2	~	9	-.03	8	-.04	80	.07
Students Orderly and Quiet								
Grade 4	84	.11	11	-.07	3	-.26	1	~
Grade 8	80	.13	11	-.08	6	-.01	2	-.38
Grade 12	89	.07	6	-.07	2	~	2	~
Students Focused on Assessment								
Grade 4	83	.12	14	-.12	2	-.35	1	~
Grade 8	78	.14	15	-.04	4	-.22	2	~
Grade 12	85	.09	12	-.13	1	~	1	~

	Very Well		Satisfactory		Unsatisfactory	
How well did the session go?	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**
Grade 4	85	.10	12	-.14	#	~
Grade 8	82	.14	14	-.16	2	~
Grade 12	82	.08	14	-.09	1	~

Debriefing form responses missing for approximately 1% of the sessions at each grade.

~ Indicates insufficient data to report achievement.

# Rounds to zero.

To allow us to compute average achievement based on all three operational assessments, students' normit scores[2] were used instead of scale scores. A common metric was necessary because NAEP reports achievement results in a different metric for civics than for U.S. history and geography.[3] Normits (also known as probits) behave somewhat like *z*-scores in that they have a mean of zero and standard deviation of 1, and they indicate where a dichotomous score or percentage lies under a normal distribution. So, a negative normit means a score or percentage below the mean, and the larger a normit is, the further it is from zero.

The average normit score at each grade would be zero for all NAEP 2010 participants, but average achievement is a bit higher than zero for students in the testing conditions study because the study includes only students in regular sessions and not those tested in separate accommodation or makeup sessions.

The first three panels in Table 2 (top of page 3) present the context for the assessment sessions by providing information about session size, location, and day of the week. As shown in the first panel, grade-by-grade average achievement in the normit metric for students in the testing conditions study was .07 for 4th grade, .08 for 8th grade, and .05 for 12th grade. To help with interpreting student achievement in the normit metric, note that the normit scale has a standard deviation of 1 compared to a standard deviation of 35 on the NAEP civics scale and of 50 on the U.S. history and geography scales. Thus, a difference of .10 normits corresponds to 3.5 points on the civics scale or 5 points on the U.S. history and geography scales.

Regarding session size, nearly all 4th-grade students were assessed in sessions of 40 or fewer. Even at grades 8 and 12, the majority of students were assessed in sessions of 40 or fewer, and achievement in these sessions was very similar to the average achievement for all students in the testing conditions study. Furthermore, although 36–37 percent of students in each of the higher grades were assessed in sessions of more than 40 students, students in these sessions did not have lower achievement than students assessed in smaller sessions.

The results by session location (second panel) show that substantial percentages of 8th-graders (59 percent) and 12th-graders (67 percent) were assessed outside of their classrooms in auditoriums, libraries, lunchrooms, or "other." "Other" session locations included multipurpose areas and nontraditional classrooms such as music or art rooms.[4] There was little variation in average achievement across session locations. The exceptions were 4th-graders tested in a library, who scored .14 normits below the average for all students in the testing conditions study, and 12th-graders tested in an auditorium, who scored .10 normits above the average.

[2] Normit scores and sampling weights were provided by Educational Testing Service, the NAEP contractor for analysis and reporting.

[3] Civics achievement is reported on a 0–300 scale for each grade, whereas U.S. history and geography are reported on 0–500 cross-grade scales.

[4] For session location and several other questions, the debriefing form provided a write-in "other" response category or space for the test administrators to provide additional details about the testing conditions. There were not many written comments, but for questions with 50 or more comments (about 5 percent of the sessions), the information was classified, and the results are in Appendix B.

Looking at the day of the week on which the sessions were conducted (third panel), at all three grades, most students were tested on Tuesday, Wednesday, or Thursday, with little variation in achievement. (Interestingly, although some of the differences were small, students assessed on Monday or Friday had higher average achievement.)

The data in the remaining panels of Table 2 address the first question of the study: Are NAEP testing conditions consistent with best assessment practices? Based on both the debriefing form questions routinely completed by Westat assessment administrators and the new questions constructed for this study, these data describe the frequency of the various problems encountered in NAEP's 2010 regular testing sessions and the levels of student achievement associated with each type of problem. Importantly, for most problem types, the results show that only small percentages of students were assessed in sessions where the assessment administrator reported problems. However, considering that average achievement was .05 to .08 for all students in the testing conditions study, it can be seen that achievement for students in problematic sessions was average or below.

Looking at the bottom panel on page 3, which includes data on the original debriefing form questions, one sees that by far the largest "problem" percentages were for sessions where some students were still working when the timer rang—69 percent of students at grade 4 and 52–56 percent at grades 8 and 12 were tested in such sessions. However, administrator comments indicated that, in sessions marked as "students still working," this often involved only 1–2 students per session (see Appendix B). Finally, average achievement in sessions with students still working when the timer rang was about the same as achievement overall, with the possible exception of 8th grade.

Looking at the results for problems more closely related to testing conditions per se, there are several areas of concern, especially for 8th- and 12th-grade students:

- Across the grades, 8–9 percent of students were in sessions with problems setting up, and at 12th grade these students scored lower on average (by .19 normits).
- For problems getting students to the sessions, the percentage of students affected increased by grade level: 10 percent at grade 4, 19 percent at grade 8, and 37 percent at grade 12. Affected students scored below average, especially at grades 4 and 8. The assessment administrators' comments indicated that some 8th- and 12th-graders were late because they were not dismissed on time from their previous classes.
- Refusals also became a problem at grade 12, with about one-fourth of students (24 percent) tested in sessions with some student refusals. However, average achievement for students in those sessions was not lower than achievement overall.
- About one-third of students at each grade (32–38 percent) were in sessions where some students left the room during testing. Again, though, achievement for these sessions was not below average. According to the administrators' comments, departures were overwhelmingly for students to go to the bathroom or get a drink.

- At grades 8 and 12, 9–10 percent of students were assessed in locations the test administrators found problematic, and the 8th-graders in such sessions had lower average achievement (by .12 normits).
- Finally, the percentages of students in sessions with interruptions increased from 10 percent at grade 4 to 16 percent at grade 8 and 19 percent at grade 12. Again, the 8th-grade students in such sessions scored lower (by .09 normits). According to the administrators' comments, the majority of the interruptions were in the form of loud noises through the intercom, phones ringing, or the occasional fire alarm.

The second page of Table 2 contains the results for the new testing conditions study questions about students being assessed in crowded, noisy, or otherwise disruptive conditions. Again, the responses indicate that the NAEP testing conditions are consistent with best assessment practices.

More specifically, the results show that assessment administrators agreed that the overwhelming majority of students had enough space to work and that there was adequate space to monitor sessions. However, small percentages of students were tested in less than optimal conditions, and those students typically had lower achievement than their peers.

- Across the three grades, 4–8 percent of students were assessed in cramped spaces, and those who were in 4th or 12th grade scored below average.
- Also, 5–7 percent were assessed in rooms with uncomfortable temperatures, and the 12th-graders among those students had lower achievement.
- At grades 8 and 12, 11–12 percent of students were assessed in noisy conditions and had lower achievement.
- Across the grades, 4–7 percent were tested in venues with visual distractions, with those who were 4th and 12th-graders scoring below average.
- At grades 8 and 12, 8–11 percent were assessed in conditions with numerous school disruptions (e.g., the PA system), although there was little relationship with average achievement.

Across the grades, 80–89 percent of the students were assessed in sessions where assessment administrators agreed "a lot" that students were quiet and orderly, and 78–85 percent were assessed in sessions where assessment administrators agreed "a lot" that students focused on the assessment. However, the 6–15 percent of students in sessions when the administrators agreed only "a little" (instead of "a lot") with either of these characterizations scored .12 to .19 normits lower than average, and the 2–8 percent of students assessed in sessions that assessment administrators characterized as disorderly or unfocused typically had very low achievement.

The last panel in Table 2 (bottom of page 4) is based on the assessment administrators' holistic judgment of how well the sessions went. According to the reports, 82–85 percent of students were tested in sessions that went "very well." Most of the rest (12–14 percent) were tested in sessions judged "satisfactory," but considering that these students scored .14 to .24 normits below average, the satisfactory sessions may not have gone all that well. Virtually no 4th-grade students and only extremely small percentages of 8th- and 12th-grade students (1–2 percent) were assessed in sessions judged to be "unsatisfactory." The comments associated with the unsatisfactory sessions indicated a range of unexpected events, with those reported multiple times having to do with student misbehavior, weather delays, having to change the location of the assessment, and running out of materials.

Number of Problems in Sessions

Although the question-by-question results indicate no single widespread problem, it is possible that students might be impacted more seriously when negative conditions pile up. Therefore, it is important to consider the percentage of affected students, and their achievement, for sessions where the assessment administrators reported multiple problems.

Table 3 shows the percentages of students and their average achievement in normits by the frequency of problems in the sessions, as measured across both the original debriefing form questions and the questions newly constructed for this study. In general, most of the students were in sessions with few problems, especially at the 4th grade. Also, essentially no students were in sessions with more than 10 problems.

Table 3. Percentage of Students and Average Achievement (in Normits) in the Testing Conditions Study of the NAEP 2010 Civics, U.S. History, and Geography Assessments by Number of Problems in Testing Session

	<2		2–5		6–10		>10	
Grade	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	50	.10	48	.05	2	-.15	#	~
Grade 8	44	.16	48	.05	7	-.03	#	~
Grade 12	35	.14	58	.02	6	-.15	#	~

Sum of "Yes" responses to yes-no questions, "Disagree a lot/Disagree a little" responses to positively worded scale questions, and "Agree a lot/Agree a little" responses to negatively worded scale questions, respectively, with missing responses deleted listwise.

~ Indicates insufficient data to report achievement. # Rounds to zero.

Because results are rounded to the nearest whole number, some totals may appear inconsistent.

The results show that achievement decreased as the number of problems in the sessions increased, so students assessed in sessions with few problems (two or fewer) generally had the highest achievement (.10 to .16 normits, on average), and students assessed in sessions with 6–10 problems had the lowest achievement (below the testing conditions study average by .11 to .22 normits). At grade 4, about half the students were in problem-free sessions (0 problems or only 1 problem), almost half were in sessions with 2–5 problems, and only about 2 percent were in sessions with 6–10 problems. The distribution did shift with each higher grade, however, to smaller percentages of students in problem-free sessions and larger percentages in sessions with multiple problems. At 8th grade, 48 percent of the students were in sessions with 2–5 problems, and 7 percent were in sessions with 6–10 problems. At 12th grade, 58 percent of students were in sessions with 2–5 problems and 6 percent were in sessions with 6–10 problems.

Analysis of the Types of Problems That Occurred in Sessions

Another way to examine the effects of testing conditions on NAEP achievement is to examine the types of problems that occurred in administering the sessions. To pursue this issue, we performed a principal components factor analysis that included both the new items developed for the study and the original questions in the session debriefing form. The idea was to summarize the data from the session debriefing form in a way that would simplify further analyses while retaining most of the information in the original questions. As shown in the top section of Table 4, the analysis of the session results, combined across all three grades, yielded three components or factors that seemed to describe the items quite well. The first factor is called *Orderly environment*, and is described by four items (adequate space for the students, adequate space for the monitors, orderly and quiet students, and focused students); the second factor is labeled *No disruptions* and is also described by four items (no interruptions, quiet room, no visual distractions, and no disruptions); and the third factor is called *No participation problems* and is defined by three items (no problems getting students to the session, no student refusals, and no students leaving the session). Three scores were built using the items that loaded on the three factors described above.

Table 4. NAEP Testing Condition Factors in the Testing Conditions Study of the NAEP 2010 Civics, U.S. History, and Geography Assessments

Factor Analysis of Problems in NAEP 2010 Testing Conditions – Factor Loadings

	Orderly Environment	No Disruptions	No Participation Problems
Adequate Space for Students to Work	.80	.13	-.12
Ample Space to Monitor Students	.70	.12	-.27
Students Orderly and Quiet	.70	.03	.50
Students Focused on Assessment	.68	.05	.52
Interruptions (R)	.01	.67	.07
Room Noisy Because of School Activity (R)	.11	.66	.08
Visual Distractions (R)	.17	.64	.13
Numerous School Disruptions (e.g., PA) (R)	.00	.78	.01
Problems Getting Students to Session (R)	-.01	.19	.41
Student Refusals (R)	-.15	.04	.59
Students Left Session (R)	.08	.02	.48

(R) Reversed

Percent of variance accounted for by three factors: 50%

Principal components analysis with varimax rotation and mean substitution – all grades together

Note: In the following panels higher factor scores denote fewer problems and lower factor scores more problems.

Average NAEP Testing Condition Factor Scores by Grade

Grade	Orderly Environment	No Disruptions	No Participation Problems
Grade 4	-.08	.19	.18
Grade 8	-.11	-.13	-.03
Grade 12	.20	-.18	-.37

Average NAEP Testing Condition Factor Scores by Session Location

Grade	Location	Orderly Environment	No Disruptions	No Participation Problems
Grade 4	Classroom	-.08	.21	.23
	Other location	-.08	.06	-.09
Grade 8	Classroom	-.11	-.01	.15
	Other location	-.11	-.21	-.16
Grade 12	Classroom	.17	-.08	-.03
	Other location	.22	-.23	-.54

Other location includes auditorium, lunchroom, library, and other.

Average NAEP Testing Condition Factor Scores by Session Size

Grade	Session Size	Orderly Environment	No Disruptions	No Participation Problems
Grade 4	<20	.17	.15	.23
	20-40	-.21	.23	.19
	41-60	-.14	.26	-.34
	>60	-.45	-.86	-.76
Grade 8	<20	.23	.03	.19
	20-40	-.18	-.07	.06
	41-60	-.07	-.23	-.29
	>60	-1.04	-.69	.65
Grade 12	<20	.28	-.26	-.28
	20-40	.19	-.15	-.28
	41-60	.18	-.13	-.60
	>60	.16	-.46	-.05

The remaining panels in Table 4 show average factor scores on the testing condition factors by grade, session location, and session size. Higher scores represent fewer problems and lower scores more problems.

The fact that the principal components analysis was conducted on the three grades combined enables interesting comparisons across the grades. At the 4th grade, having disorderly sessions was more of an issue than having problems with disruptions or with participation, whereas the opposite was true at grade 12. In general, the 8th grade had few problems with participation, but there were some problems with disorderly environment and disruptions.

At all three grades, sessions conducted outside of traditional classrooms were more prone to disruptions and problems with participation (although at grade 4, most sessions were in classrooms). At all grades, compared to classroom sessions, those in other locations had lower factor scores due to disruptions, and, especially, due to problems with participation (-.54 at 12th grade).

Analysis of Testing Conditions by Student Groups

For the most part, NAEP 2010 testing conditions were consistent with best practices, but assessment administrators reported that some students were assessed in less than optimal conditions. Also, the results show that less than optimal testing conditions were associated with lower achievement. For example, students assessed in sessions with multiple problems had lower achievement. Also, when testing condition problems were grouped into factor scores, students were found to have lower achievement if they were tested in sessions characterized as disorderly, with disruptions, or with participation problems.

It may be, however, that below-standard testing conditions are related to school poverty. Because schools in economically depressed areas with large percentages of minority and economically disadvantaged students may be among those schools most likely to have overcrowded and noisy conditions, they may be particularly prone to having disorderly and disruptive NAEP testing sessions. If poor testing conditions primarily reflect the poor schooling conditions for students' everyday learning experiences, then lower performance may reflect schooling conditions more than testing conditions. If not, however, differences in testing conditions could be impacting the achievement gaps routinely reported by NAEP.

NAEP routinely reports differences in average achievement among students in five race/ethnicity groups, among levels of eligibility for the national school lunch program, and between English language learners (ELLs) and non-English language learners (non-ELLs). Table 5 contains the percentage and average achievement (in normits) of students in the testing conditions study for each of NAEP's reporting categories of race/ethnicity, eligibility for the national school lunch program, and ELLs, by grade level. Although the testing conditions study results are based on the combined total of students assessed in civics, U.S. history, and geography, and do not include students accommodated outside of the regular sessions or assessed in makeup sessions, the percentages and gaps in average achievement are relatively consistent with those reported in the 2010 *Report Cards* for these subjects.

In the testing conditions study, across the grades, the percentages for the five racial/ethnic groups were 58–62 percent White students, 13–15 percent Black students, 16–20 percent Hispanic students, 5–7 percent Asian/Pacific Islander students, and 1 percent American Indian/Alaskan Native students. White students scored .76 to .82 normits higher on average than Black students and .58 to .74 normits higher than Hispanic students.

NAEP uses eligibility for the national school lunch program as an indicator of low income since students from lower-income families are eligible for either free or reduced-price school lunches. In the testing conditions study, 44 percent of the 4th-graders, 40 percent of the 8th-graders, and 30 percent of the 12th-graders were eligible for a free or reduced-price lunch. The eligible students scored .61 to .77 normits lower on average than students not eligible for the school lunch program.

Compared to the 93–97 percent of students who were not classified as ELLs, the small percentage of ELLs (3–7 percent) assessed in regular sessions performed about 1 normit (one standard deviation) lower on average.

Table 5. Percentage and Average Achievement (in Normits) in the Testing Conditions Study of the NAEP 2010 Civics, U.S. History, and Geography Assessments by NAEP Reporting Categories

	Race/Ethnicity									
	White		Black		Hispanic		Asian/ Pacific Islander		American Indian/ Alaskan Native	
	Percent Students	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**	**Percent Students**	**Average Achieve.**
Grade 4	58	.34	15	-.42	20	-.40	5	.33	1	-.30
Grade 8	60	.33	14	-.44	19	-.35	5	.31	1	-.22
Grade 12	62	.26	13	-.56	16	-.32	7	.16	1	-.25

	National School Lunch Program					
	Eligible		Not Eligible		Information Not Available	
	Percent Students	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**
Grade 4	44	-.37	49	.40	7	.48
Grade 8	40	-.34	54	.36	6	.46
Grade 12	30	-.39	63	.22	7	.31

	English Language Learners			
	Yes		No	
	Percent Students	**Average Achievement**	**Percent Students**	**Average Achievement**
Grade 4	7	-.81	93	.14
Grade 8	4	-.92	96	.13
Grade 12	3	-1.04	97	.08

Are Disadvantaged Students More Likely to Be Found in Sessions With Certain Types of Problems? In Sessions With High Numbers of Problems?

As mentioned above, we are interested in both the number of problems that occur in a session and in the types of problems. Here we examine whether disadvantaged students are more likely to be in sessions with a large number of problems or in sessions with one of the three types of problems measured by the three factor scores. Table 6 examines average testing condition factor scores, and the percentages of students assessed in sessions with six or more problems, by NAEP reporting category. The results show the extent to which minority and economically disadvantaged students were assessed in more problem-prone testing conditions than their peers. For each of the testing condition factors, a lower score indicates being in a session with more problems. Of course, it is important to remember that students are assessed in the schools that they attend. With survey data it is impossible to determine the degree to which the NAEP testing conditions are primarily a reflection of school conditions or specific to the sessions themselves.[5]

Table 6. Types of Problems in Testing Conditions (Average Factor Scores) in the Testing Conditions Study of the NAEP 2010 Civics, U.S. History, and Geography Assessments by Selected NAEP Reporting Categories

	NAEP Testing Conditions Average Factor Scores			Percent of Students Assessed in Sessions with 6 or More Problems
	Orderly Environment	No Disruptions	No Participation Problems	
Grade 4				
White	-.03	.17	.21	2
Black	-.21	.24	.15	5
Hispanic	-.12	.22	.09	1
Other*	-.09	.19	.19	2
Eligible Free and Reduced Lunch	-.15	.20	.12	3
Not Eligible	-.05	.17	.20	2
English Language Learner	-.22	.19	.00	2
Not English Language Learner	-.07	.19	.19	2
Grade 8				
White	-.09	-.08	.03	7
Black	-.17	-.20	-.07	9
Hispanic	-.17	-.16	-.17	8
Other*	-.05	-.28	-.03	7
Eligible Free and Reduced Lunch	-.15	-.17	-.11	9
Not Eligible	-.09	-.12	.00	7
English Language Learner	-.32	-.28	-.16	10
Not English Language Learner	-.10	-.12	-.03	7
Grade 12				
White	.23	-.15	-.34	6
Black	.17	-.27	-.55	9
Hispanic	.13	-.23	-.36	10
Other*	.20	-.23	-.27	6
Eligible Free and Reduced Lunch	.14	-.25	-.38	9
Not Eligible	.27	-.17	-.43	6
English Language Learner	.11	-.27	-.16	9
Not English Language Learner	.21	-.18	-.37	7

* Other includes Asian/Pacific Islander, American Indian/Alaskan Native, and other.

At 4th grade, Black and Hispanic students, students eligible for free or reduced-price lunch, and ELLs had lower scores on two out of the three testing condition factors—*Orderly environment* and *No participation problems*—than White students, those

[5] We will examine this question further in a set of analyses described in the subsequent sections on HLM analysis.

not eligible for the lunch program, and non-ELLs, respectively. Also, 5 percent of the Black students were assessed in sessions with six or more problems, compared to 1 to 2 percent of the Hispanic students or those of another race/ethnicity.

At 8th grade, Black and Hispanic students, students eligible for free or reduced-price lunch, and ELLs had lower average scores on all three testing condition factors than their more advantaged peers. Also, slightly higher percentages (1–2 percent) of Black and Hispanic students than White students were assessed in sessions with six or more problems. The same was true for students eligible for free or reduced-price lunch compared to those not eligible. Finally, 10 percent of the ELLs, compared to 7 percent of the non-ELLs, were assessed in sessions with six or more problems.

At 12th grade, Black students had lower average scores than White students on the *No disruptions* and *No participation problems* factors, while Hispanic students had lower average scores on the *Orderly environment* and *No disruptions* factors. Students eligible for free or reduced-price lunch and ELLs also had lower factor scores on *Orderly environment* and *No disruptions.* Higher percentages of Black and Hispanic students than White students (9–10 percent compared to 6 percent) were assessed in sessions with six or more problems. Similarly, 9 percent of students eligible for free or reduced-price lunch and 9 percent of ELLs were assessed in sessions with six or more problems, compared, respectively, with 6 percent of students not eligible for the lunch program and 7 percent of non-ELLs.

In summary, there is some evidence, though not particularly strong, that at all three grade levels, students disadvantaged by minority status, poverty, or ELL status are more likely to be in testing sessions with a higher number of problems and more likely to be in sessions that have worse factor scores on at least two of the three factors describing types of testing conditions problems.

Using HLM Analysis to Examine the Effects of the Number and Types of Session Problems on NAEP Civics Scores

In this section, we examine the effects of the number of problems as well as the effects of the types of problems on NAEP achievement scores in civics at grades 4, 8, and 12. We then ask whether any statistically significant relationships between session problems and achievement persist when we control for (1) individual student socio-demographic variables (race/ethnicity, poverty, and ELL status) and (2) the socio-demographic status of the schools the students are attending. These relationships were examined using HLM analyses.

So that the HLM analyses would be in the NAEP scale score metric, the analyses were conducted using plausible values. However, as noted earlier, for civics achievement, NAEP uses a 0–300 scale for each grade, whereas U.S. history and geography are reported on 0–500 cross-grade scales. The different scales meant that the subjects could not be combined and remain in their respective metrics. Therefore, the HLM analyses were conducted using only the civics scale scores and session data. Because NAEP calculates five plausible values for each student, the analyses were conducted using HLM software (Raudenbush, Bryk, & Congdon,

2004) that analyzed each plausible value in turn and then averaged across the five analyses as the final result.

The civics HLM analyses were based on approximately one-third of the students in the testing conditions study. More specifically, the civics data for the testing conditions study consisted of 738 sessions with 5,554 students at grade 4, 468 sessions with 7,917 students at grade 8, and 391 sessions with 7,799 students at grade 12.[6] (See Appendix C, which also provides the sample sizes for U.S. history and geography.) Although there were some variations between the civics sessions and the U.S. history/geography sessions, the session debriefing form responses for the civics sessions were similar to those overall (see Appendix D for session debriefing form responses analyzed separately for the three subject areas).

Testing conditions were characterized in two ways in the HLM analyses. To give a global indication of the quality of the testing conditions, each session was categorized according to the number of problems reported by the assessment administrator—fewer than 2, 2–5, and 6 or more. Each session also was assigned a score on each of the three factors measuring testing conditions—*Orderly environment*, *No disruptions*, and *No participation problems*. (Since these are essentially two ways of conceptualizing the same data—administrators' reports of problems in sessions—one cannot include both in a single statistical model.)

The analysis included three individual student indicators of socio-demographic status: whether or not the student was Black or Hispanic, whether or not the student was eligible for free or reduced-price lunch, and whether or not the student was categorized by the school as ELL. In addition to the individual socio-demographic indicators, two indicators of school poverty were included in the analyses: whether or not more than 50 percent of students in the school were Black or Hispanic, and whether or not more than 50 percent of students were eligible for free or reduced-price lunch.

Because the conditions of testing analyses relate student achievement and indicators of disadvantage to characteristics of the sessions in which the students were grouped for test administration purposes, it was natural to use a hierarchical linear modeling (HLM) approach, with students as a first level nested within sessions as a second level. For these analyses, a school's poverty indicators were considered to apply to the testing sessions conducted in that school, and accordingly were treated as session-level variables.

At each grade, relationships between testing conditions and student achievement were analyzed first according to number of problems reported and second in terms of the three testing condition factors. For each of the two characterizations of testing conditions, three models were constructed showing student achievement in relation

[6] Also, because the testing conditions study included only students tested in regular sessions (not in separate accommodated sessions or makeup sessions), there were fewer students than for the results reported in *The Nation's Report Card: Civics 2010* (National Center for Education Statistics, 2011): about 1,600 less at grades 4 and 8, and about 2,100 less at grade 12.

to (1) testing conditions alone, (2) testing conditions and individual student indicators of disadvantage, and (3) testing conditions, individual student indicators of disadvantage, and school socio-demographic indicators.

Grade 4 Results

Table 7 summarizes the results of the 4th-grade analyses, with Models 1 through 3 representing testing conditions in terms of number of problems reported in the session, and Models 4 through 6 representing testing conditions in terms of scores on the three testing condition factors.

Table 7. Grade 4 HLM Models Showing Relationship Between Testing Conditions and Average Score in the NAEP 2010 Civics Assessment for Students in the Testing Conditions Study

	Number of Problems in the Sessions			Testing Condition Factors		
Grade 4	**Model 1 Without Controls**	**Model 2 With Student Controls**	**Model 3 With Student and School Controls**	**Model 4 Without Controls**	**Model 5 With Student Controls**	**Model 6 With Student and School Controls**
Intercept Predicted Mean Civics Score	161.6 (1.0)	172.2 (0.9)	174.5 (1.0)	160.7 (0.8)	172.0 (0.8)	174.5 (0.8)
Number of Problems (Reference Category <2 Problems)						
2–5 Problems	-1.4 (1.6)	-0.1 (1.1)	0.3 (1.1)			
6 or More Problems	-11.9** (5.2)	-3.5 (3.6)	-0.7 (3.2)			
Testing Condition Factors						
Orderly Environment				1.3* (0.7)	0.7 (0.5)	0.6 (0.5)
No Disruptions				0.7 (1.0)	0.2 (0.7)	-0.02 (0.6)
No Participation Problems				3.5*** (0.9)	1.8*** (0.6)	1.3** (0.6)
School Demographics						
More than 50% of Students Black or Hispanic			-4.8*** (1.3)			-4.6*** (1.3)
More than 50% of Students Eligible for Free or Reduced Lunch			-7.3*** (1.2)			-7.2*** (1.2)
Student-Level Variables						
Black or Hispanic		-10.4*** (1.0)	-7.7*** (1.1)		-10.3*** (1.0)	-7.7*** (1.1)
Eligible for Free or Reduced Lunch		-14.6*** (1.0)	-12.4*** (1.0)		-14.5*** (1.0)	-12.3*** (1.0)
English Language Learner		-20.2*** (1.7)	-19.0*** (1.7)		-20.1*** (1.7)	-19.0*** (1.7)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.

****p* < .01. ***p* < .05. **p* <.10.

Standard errors for model parameters are shown in parentheses.

Effects of the Number of Problems Reported. In Models 1 through 3, number of problems is coded so that sessions that were problem free (fewer than 2 problems) are taken as a reference category, and the other two categories, 2–5 problems and 6 problems or more, as deviations from this reference category. Accordingly, the predicted mean NAEP civics score shown in Model 1 (161.6) is the mean achievement estimated by this model for 4th-grade students in problem-free sessions. The model estimates that students in sessions with 2–5 problems would have slightly lower achievement (by 1.4 points), although this difference is not statistically significant. However, 4th-grade students in sessions with 6 or more problems could be expected to have average achievement almost 12 points lower than students in problem-free sessions. (As a basis of comparison, for the 2010 civics assessment, the difference between achievement at the 90th and 75th percentiles was 15 points.)

Model 2 is an extension of Model 1 that controls for three student-level variables: being Black or Hispanic, being eligible for free or reduced-price lunch, and being a student categorized as ELL. For Model 2, the predicted mean civics score (172.2) is the model's average achievement estimate for 4th-grade students tested in problem-free sessions who are not Black or Hispanic, not eligible for free or reduced-price lunch, and not ELLs. Not surprisingly, this expected score is considerably greater than the mean predicted by Model 1. Model 2 clearly shows the expected disadvantage associated with the three student indicators. According to this model, students tested in problem-free sessions who are Black or Hispanic could be expected to score lower by 10.4 points, on average, than other students tested in such sessions. Similarly, students tested in problem-free sessions who are eligible for free or reduced-price lunch could be expected to score lower by 14.6 points, and those who are ELLs could be expected to score lower by 20.2 points.

The most noteworthy aspect of Model 2 compared with Model 1 is the change in the expected effect of being tested in a session with 6 problems or more. The adverse effect of being tested in such a session drops from 11.9 points for students in general to 3.5 points (which is not statistically significant) after controlling for student race/ethnicity and poverty.

Finally, the most important analysis, shown in Model 3, examines the effects of the number of problems in the sessions, taking into account both student- and school-level variables. Here, the adverse effect of being tested in a session with many problems is reduced to 0.7 points when controlling for school poverty in addition to individual disadvantage. That is, the effects of the number of problems in the session washes out when taking into account the student-level variables and the school's percentage of minority students as well as its poverty level.

Effects of the Three Categories of Testing Conditions. Models 4 through 6 in Table 7 represent the same analyses as Models 1–3, except with the testing condition factor scores instead of number of problems reported. A further difference is that the factor scores are continuous variables, whereas number of problems was represented as a single categorical variable. The predicted mean civics score shown in Model 4 (160.7) is the mean achievement estimated by the model for the 4th-grade students in the analysis. The factor score effects show how much this predicted mean would be expected to change for each one-standard-deviation change in the factor scores. Two of the three factors, *Orderly environment* and *No participation problems*, have statistically significant effects (1.3 and 3.5 points, respectively). According to Model 4, if problems with participation could be reduced to the extent that the average factor score increased by one standard deviation, then average achievement could be expected to rise by 3.5 points.

Model 5 shows that controlling for individual student disadvantage reduces the effect of the *Orderly environment* factor from 1.3 to 0.7 points (no longer statistically significant). The effect of the *No participation problems* factor is reduced from 3.5 to 1.8 points, which is still statistically significant. Controlling for school poverty and the percentage of minority students in the school further reduces that effect to 1.3 points (Model 6). Although the effect is small, it remains statistically significant.

To summarize, at grade 4, the number of problems that occurred in a session is not statistically significant when student and school-level socio-demographics are controlled. When the three testing condition factors are considered, the effects of these variables also are washed out when student- and school-level socio-demographics are controlled, except for the effect of *No participation problems* (where students arrive on time, do not refuse to participate, and stay through the session).

Grade 8 Results

Table 8 (Models 1–6) presents the results of an identical series of analyses for the 8th grade.

Table 8. Grade 8 HLM Models Showing Relationship Between Testing Conditions and Average Score in the NAEP 2010 Civics Assessment for Students in the Testing Conditions Study

	Number of Problems in the Sessions						Testing Condition Factors					
Grade 8	Model 1 Without Controls		Model 2 With Student Controls		Model 3 With Student and School Controls		Model 4 Without Controls		Model 5 With Student Controls		Model 6 With Student and School Controls	
Intercept Predicted Mean Civics Score	158.2	(1.1)	166.7	(1.0)	168.9	(1.0)	155.1	(0.9)	165.4	(0.8)	167.9	(0.8)
Number of Problems (Reference Category <2 Problems)												
2–5 Problems	-6.2***	(1.8)	-2.4*	(1.3)	-1.5	(1.3)						
6 or More Problems	-8.3	(5.4)	-4.0	(3.4)	-3.8	(3.0)						
Testing Condition Factors												
Orderly Environment							1.4	(0.9)	1.1*	(0.6)	1.2**	(0.6)
No Disruptions							2.7***	(1.0)	1.3**	(0.6)	0.9*	(0.5)
No Participation Problems							3.9***	(1.0)	2.4***	(0.7)	2.0***	(0.7)
School Demographics												
More than 50% of Students Black or Hispanic					-2.4	(1.5)					-2.0	(1.5)
More than 50% of Students Eligible for Free or Reduced Lunch					-8.3***	(1.3)					-8.1***	(1.3)
Student-Level Variables												
Black or Hispanic			-10.3***	(1.0)	-9.1***	(1.0)			-10.2***	(1.0)	-9.0***	(1.0)
Eligible for Free or Reduced Lunch			-13.4***	(0.9)	-12.1***	(0.9)			-13.3***	(0.9)	-12.1***	(0.9)
English Language Learner			-35.5***	(2.1)	-34.6***	(2.2)			-35.5***	(2.1)	-34.6***	(2.2)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.
***p < .01. **p < .05. *p <.10.
Standard errors for model parameters are shown in parentheses.

Effects of the Number of Problems Reported. When considering the number of problems in a session, the results for grade 8 are very similar to those for grade 4. Any effects found with no controls, or controlled only for student-level variables, no longer are significant when school level socio-demographics are also included in the model. That is, school-level factors seem to account for the effects observed rather than the number of problems in a session.

Effects of the Three Categories of Testing Conditions. By way of contrast, all three of the testing condition indicators are significantly associated with 8th-grade civics scores even when student- and school-level socio-demographic factors are taken into account. The *No participation problems* factor has the greatest effect—3.9 NAEP points before any statistical adjustments, reduced to 2 points after controlling for student and school socio-demographic factors. The *No disruptions* factor has an effect before adjustment of 2.7 NAEP points, and is reduced to 0.9 points after controlling for student and school socio-demographic factors. The *Orderly environment* factor effect is reduced from 1.4 NAEP points before adjustment to 1.2 points after student and school controls are applied.

To summarize, at grade 8, as was the case for grade 4, when individual- and school-level factors are controlled, the number of problems in a session is unrelated to NAEP civics performance. By contrast, the effects of each of the three types of session problem remain statistically significant, although the effects are not large.

Grade 12 Results

Table 9 (Models 1–6) shows the results of the analyses for 12th grade.

Table 9. Grade 12 HLM Models Showing Relationship Between Testing Conditions and Average Score in the NAEP 2010 Civics Assessment for Students in the Testing Conditions Study

	Number of Problems in the Sessions			Testing Condition Factors		
Grade 12	Model 1 Without Controls	Model 2 With Student Controls	Model 3 With Student and School Controls	Model 4 Without Controls	Model 5 With Student Controls	Model 6 With Student and School Controls
Intercept Predicted Mean Civics Score	151.7 (1.3)	158.3 (1.2)	160.3 (1.1)	149.8 (0.9)	157.8 (0.9)	159.8 (0.9)
Number of Problems (Reference Category <2 Problems)						
2–5 Problems	-2.7 (1.9)	0.1 (1.5)	0.1 (1.5)			
6 or More Problems	-10.8** (5.1)	-8.2* (4.3)	-7.8* (4.1)			
Testing Condition Factors						
Orderly Environment				4.6*** (1.3)	3.5*** (1.1)	3.2*** (1.1)
No Disruptions				2.2*** (0.8)	1.5** (0.6)	1.3** (0.6)
No Participation Problems				1.3 (0.9)	1.3* (0.8)	1.5** (0.7)
School Demographics						
More than 50% of Students Black or Hispanic			0.6 (2.0)			1.5 (1.9)
More than 50% of Students Eligible for Free or Reduced Lunch			-10.1*** (1.9)			-10.2*** (1.9)
Student-Level Variables						
Black or Hispanic		-12.8*** (1.1)	-11.9*** (1.1)		-12.7*** (1.1)	-11.9*** (1.1)
Eligible for Free or Reduced Lunch		-10.8*** (1.1)	-9.9*** (1.2)		-10.7*** (1.1)	-9.9*** (1.1)
English Language Learner		-37.5*** (2.5)	-37.1*** (2.5)		-37.5*** (2.5)	-37.0*** (2.5)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.
***p < .01. **p < .05. *p <.10
Standard errors for model parameters are shown in parentheses.

Effects of the Number of Problems Reported. Unlike in grades 4 and 8, when individual and school level socio-demographics are included in the model for grade 12, there remains a large, statistically significant effect of being in a session with 6 or more problems. On average for students in such a session, NAEP civics scores are nearly 8 point lower (compared to a nearly 11-point difference without individual- and school-level controls).

Effects of the Three Categories of Testing Conditions. Among the testing condition factors, being in a session with no problems with orderliness (the *Orderly environment* factor) shows the strongest relationship to NAEP achievement scores when individual- and school-level controls are included (a 3.2-point difference), followed by being in a session with *No participation problems* (a 1.5-point difference), and finally in a session with *No disruptions* (a 1.3-point difference). All three of these testing condition category effects are statistically significant, but quite small.

Summary

In summary, assessing students in crowded, noisy, or otherwise disruptive conditions is associated with lower performance, but there may be overriding factors involved. The HLM analyses found that few of the zero-order relationships between the number of problems in a session and NAEP civics achievement hold up when individual level and school level socio-demographics are controlled. The one exception was at grade 12, where being in a session with 6 or more problems was substantially related to NAEP civics scores. By way of contrast, the types of problems seemed to matter more than the number of problems. Being in a session where there were no participation problems was statistically significant in all three grade levels even after individual and school demographics were controlled. Being in an orderly session or a session with few or no disruptions also was significant at grades 8 and 12. Being in an orderly session was especially important at grade 12.

The HLM Analysis of Racial/Ethnic Achievement Gaps in Civics

Noting that minority and economically disadvantaged students were more likely to be assessed under poor testing conditions, and that being assessed under poor testing conditions appears to have an adverse effect on achievement beyond attending a disadvantaged school, we undertook further HLM analyses to inform NAEP's reporting of achievement gaps. More specifically, we undertook to present achievement estimates for the disadvantaged groups first, and then show how the estimates change once they are controlled for testing conditions.

Two sets of HLM analyses to examine the White–Black gap and the White–Hispanic gap in relation to testing conditions were conducted. Consistent with the previous HLM analyses, testing condition indicators included the number of problems reported by the assessment administrator and the three factor scores—*Orderly environment*, *No disruptions*, and *No participation problems*. Because only the White and Black students were included in the White–Black gap analyses and only the White and Hispanic students were included in the White–Hispanic gap analysis, the sample sizes were smaller than for the first set of HLM analyses based on NAEP civics scores (by about one-fourth for the White–Black gap analyses and one-fifth for the

White–Hispanic gap analyses) and, thus, effects needed to be somewhat larger to be significant. The sample sizes for the HLM gap analyses are shown in Table 10.

Table 10. Number of Students in HLM Analyses of Relationship Between Racial/Ethnic Achievement Gaps and Testing Conditions in the NAEP 2010 Civics Assessment

White–Black Achievement Gap			
	Number of Students	Percent White Students	Percent Black Students
Grade 4	3,823	79	21
Grade 8	5,562	81	19
Grade 12	5,596	84	16

White–Hispanic Achievement Gap			
	Number of Students	Percent White Students	Percent Hispanic Students
Grade 4	4,063	75	25
Grade 8	5,951	76	24
Grade 12	5,861	80	20

Six gaps in average civics scores were examined, including the White–Black gap at three grades and the White–Hispanic gap at three grades. At each grade, the gap analysis began with a simple regression to match as closely as possible the NAEP results in the *Report Card: Civics 2010* (National Center for Education Statistics, 2011). The average achievement estimates are about 2 scale points higher than those reported by NAEP for the racial/ethnic groups, and the estimated gaps are very close but slightly smaller (1–2 percentage points). Presumably the differences from official NAEP reported scores are because the testing conditions data do not include students accommodated outside of regular sessions and assessed in makeup sessions.

The HLM gap analysis at each grade included 10 models—2 to estimate the student and school effects on the gap, 3 to estimate the effects of the number of problems reported by the assessment administrator, and 3 to estimate the effects of the three testing condition factors. For each gap analysis, Model 1 predicts the effect on mean civics achievement of the student being either Black or Hispanic controlling for between-school variance,[7] and Model 2 extends the first model by also controlling for the minority school indicator, which identifies schools where more than 50 percent of the students in the school are Black or Hispanic. Model 3 predicts the effect of testing conditions alone (number of problems), and Model 4 extends Model 3 by controlling for students being either Black or Hispanic and the session being conducted in a minority school. Model 5 extends Model 4 by also controlling for the interaction between number of problems and sessions being conducted in minority

[7] The between-school variance for the White–Black gap analyses was 26 percent at grade 4, 26 percent at grade 8, and 19 percent at grade 12. For the White–Hispanic gap analyses, it was 30 percent at grade 4, 24 percent at grade 8, and 15 percent at grade 12.

schools. Models 6–8 are the same as Models 3–5, but testing conditions are analyzed here in terms of the three testing condition factors.

White–Black Gap Analysis

Grade 4 Analyses. Table 11 presents the results of the 4th-grade HLM analysis of the relationship between testing conditions and the White–Black gap in average civics scores. For the race/ethnicity variable, White students are the reference group, and for the minority school variable, students not tested in a session in a minority school are the reference group. The simple regression estimates White students' average civics achievement to be 169 scale score points and that Black students would have significantly lower achievement, by 23.0 points. NAEP's *Report Card: Civics 2010* (National Center for Education Statistics, 2011) reported a 24-point score gap. (White students scored 167 points on average, compared to 143 points for Black students.)

Table 11. Grade 4 HLM Models Showing Relationship Between White–Black Achievement Gap in the NAEP 2010 Civics Assessment and Testing Conditions for Students in the Testing Conditions Study

				Number of Problems in Sessions			Testing Condition Factors		
Grade 4	**Single Level Regression Model**	**Model 1 Student Level Gap**	**Model 2 Student Level Gap With School Control**	**Model 3 Without Controls**	**Model 4 Student Level Gap With School and Testing Conditions Controls**	**Model 5 With Student, School, Testing Conditions, and Testing Conditions × School Controls**	**Model 6 Without Controls**	**Model 7 Student Level Gap With School and Testing Conditions Controls**	**Model 8 With Student, School, Testing Conditions, and Testing Conditions × School Controls**
Intercept Predicted Mean Civics Score	169 (0.8)	168.1 (0.8)	169.5 (0.8)	164.6 (1.1)	169.9 (1.0)	169.8 (1.1)	168.1 (0.8)	169.3 (0.8)	169.4 (0.8)
Student-Level Variable									
Black	-23.0*** (1.4)	-19.6*** (1.4)	-15.6*** (1.6)		-15.6*** (1.6)	-15.6*** (1.6)		-15.6*** (1.6)	-15.5*** (1.6)
Minority School									
More Than 50% of Students Black or Hispanic			-11.0*** (1.7)		-10.8*** (1.7)	-10.3*** (2.4)		-10.4*** (1.7)	-10.2*** (1.7)
Number of Problems (Reference Category <2 Problems)									
2–5 Problems				-1.3 (1.6)	-0.8 (1.4)	-0.7 (1.6)			
6 or More Problems				-11.3** (5.4)	-4.1 (3.5)	0.0 (4.0)			
Interaction Between Number of Problems and Minority School									
2-5 Problems × More Than 50% of Students Black or Hispanic						-0.1 (3.0)			
6 or More Problems × More Than 50% of Students Black or Hispanic						-9.0 (7.1)			
Testing Condition Factors (reverse coded)									
Orderly Environment							-1.3* (0.8)	-0.6 (0.6)	-0.3 (0.7)
No Disruptions							-1.0 (1.1)	-1.3 (0.9)	-1.3 (1.0)
No Participation Problems							-3.6*** (0.9)	-2.1*** (0.8)	-1.8* (1.0)
Interaction Between Testing Condition Factors and Minority School									
Orderly Environment × More Than 50% of Students Black or Hispanic									-1.0 (1.5)
No Disruptions × More Than 50% of Students Black or Hispanic									0.3 (1.8)
No Participation Problems × More Than 50% of Students Black or Hispanic									-0.7 (1.8)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.

****p* < .01. ***p* < .05. **p* <.10

Standard errors for model parameters are shown in parenthesis.

According to Model 1 of the HLM analysis, with students nested within sessions as a second level, Black students are estimated to score 19.6 points lower than White students, a smaller gap than estimated by the single level regression because the school-level variance has been removed (approximately 27 percent of the variance at 4th grade).

In Model 2, the effect of being tested in a minority school was -11 points. In the presence of the minority school variable, the estimated average achievement gap for Black students is further reduced to 15.6 points, so a Black student tested in a minority school would have a predicted score that was 26.6 points lower than a White student tested in a nonminority school.

In Models 3–5 for 4th grade, the reference category is problem-free sessions (fewer than 2 problems). In Model 3, predicting changes in achievement only according to the number of problems reported by the assessment administrator, students (White and Black) in problem-free sessions have estimated mean civics achievement of 164.6, while those in sessions with 6 or more problems could be expected to score 11.3 points lower. This large testing-conditions effect is consistent with the first set of HLM analyses.

Model 4 extends the analysis of Model 3 by adding controls for the students being Black and the session being conducted in a minority school. The estimated White–Black gap is approximately the same as in Model 2, but the negative effect found in Model 3 of being assessed in a session with 6 or more problems is reduced to 4.1 points and is no longer significant.

Since a disproportionate percentage of Black students were tested in minority schools, Model 5 extends Model 4 by controlling for the interaction between number of problems and sessions in minority schools. Compared to Model 4, the negative effect of being tested in a session with 6 or more problems is eliminated (reduced to 0.0). Also, the interaction effect associated with being assessed in a session with 6 or more problems that was conducted in a minority school is not significant.

Models 6–8 represent the same analyses, except with the testing condition factors instead of number of problems reported by the assessment administrator. Because the factor scores are continuous variables, the effects in Model 6 show how much the predicted mean civics score of 168.1 (for 4th-grade students who are White or Black) would be expected to change for each one-standard-deviation change in the factor scores. In Model 6, similar to the initial 4th-grade HLM analysis at grade 4, the factor effects are significant for *Orderly environment* and *No participation problems.*

In Model 7, extending Model 6 controlling for the student being Black and the session being held in a minority school, the estimated size of the White–Black gap remains about the same as in Model 2 (identical effect for student being Black and only slightly smaller effect for being tested in a minority school), and the *No participation problems* factor effect remains significant (reduced from -3.6 to -2.1).

In Model 8, extending Model 7 by controlling for the interaction with sessions in minority schools, there still is a small but significant effect for the factor *No* *participation problems*. Also, there is essentially no interaction between the testing condition factors and sessions in minority schools.

Grade 8 Analyses. Table 12 presents the 8th-grade HLM analyses of testing conditions in relation to the White–Black achievement gap. The simple regression model shows a predicted mean civics score for White students of 163 and predicted achievement for Black students that is lower by 23.8 points. NAEP reported a 25-point gap between White students' 2010 average civics achievement (160 points) and that of Black students (135 points).

Model 2 estimates that students tested in minority schools would score 10.6 points lower than those tested in nonminority schools, a result similar to that for 4th grade. Controlling for the session being held in a minority school reduces the estimated difference between Black and White students to 16.5 points. In Model 3, there is a significant effect for 2–5 problems, but the effect for 6 or more problems is not significant (most likely due to the small sample size).

In Models 4 and 5, controlling for number of problems, the effect of the student being Black remains the same. But in Model 5 the effect of being tested in a minority school is reduced somewhat and there is a significant 20.9-point negative effect associated with the interaction between sessions with 6 or more problems and sessions being conducted in minority schools. As a point of reference, at the 8th grade the difference between the 50th and 75th percentiles in 2010 NAEP civics scores was 20 points.

Table 12. Grade 8 HLM Models Showing Relationship Between White - Black Achievement Gap in the NAEP 2010 Civics Assessment and Testing Conditions for Students in the Testing Conditions Study

				Number of Problems in Sessions			Testing Condition Factors		
Grade 8	**Single Level Regression Model**	**Model 1 Student Level Gap**	**Model 2 Student Level Gap With School Control**	**Model 3 Without Controls**	**Model 4 Student Level Gap With School and Testing Conditions Controls**	**Model 5 With Student, School, Testing Conditions, and Testing Conditions × School Controls**	**Model 6 Without Controls**	**Model 7 Student Level Gap With School and Testing Conditions Controls**	**Model 8 With Student, School, Testing Conditions, and Testing Conditions × School Controls**
Intercept Predicted Mean Civics Score	163 (0.9)	162.1 (0.8)	164 (0.9)	160.3 (1.1)	164 (1.1)	164.2 (1.1)	158.0 (0.9)	163.5 (0.9)	164 (0.9)
Student-Level Variable									
Black	-23.8*** (1.8)	-19.1*** (1.4)	-16.5*** (1.4)		-16.4*** (1.4)	-16.5*** (1.4)		-16.4*** (1.4)	-16.4*** (1.4)
Minority School									
More Than 50% of Students Black or Hispanic			-10.6*** (2.0)		-10.4*** (2.0)	-8.8*** 3.0		-9.8*** (1.9)	-9.5*** (2.0)
Number of Problems (Reference Category <2 Problems)									
2–5 Problems				-4.7** (1.8)	-1.4 (1.6)	-1.3 (1.8)			
6 or More Problems				-6.0 (5.2)	-3.5 (3.9)	0.8 (3.5)			
Interaction Between Number of Problems and Minority School									
2–5 Problems × More Than 50% of Students Black or Hispanic						-1.1 (3.9)			
6 or More Problems × More Than 50% of Students Black or Hispanic						-20.9*** (7.8)			
Testing Condition Factors (reverse coded)									
Orderly Environment							-1.7* (0.9)	-1.4** (0.7)	-0.9 (0.8)
No Disruptions							-1.7** (0.8)	-0.7 (0.7)	-0.6 (0.8)
No Participation Problems							-3.5*** (1.0)	-2.5*** (0.8)	-2.0** (1.0)
Interaction Between Testing Condition Factors and Minority School									
Orderly Environment × More Than 50% of Students Black or Hispanic									-1.9 (1.5)
No Disruptions × More Than 50% of Students Black or Hispanic									-0.2 (1.7)
No Participation Problems × More Than 50% of Students Black or Hispanic									-1.2 (1.7)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.

***p < .01. **p < .05. *p <.10

Standard errors for model parameters are shown in parenthesis.

Looking at the 8th-grade results for the testing condition factors, all three factors are significant when considered in isolation (Model 6). Comparing Model 7 to Model 2, the effect for the student being Black remains nearly identical while the effect for the testing session being in a minority school remains similar (reduced to 9.8 from 10.6 in Model 2). Also, the negative effects remain significant for the *Orderly environment* and *No participation problems* factors. In Model 8, the negative effect of participation problems is still significant (but reduced to 2.0 points). Similar to the 4th grade, there is no significant interaction effect between the factor scores and sessions being held in minority schools.

Grade 12 Analyses. Table 13 presents the 12th-grade results for the HLM analyses of test conditions in relation to the White–Black gap. The single-level regression estimated White students' 2010 average civics achievement (157.1 points) to be 27.3 points higher than Black students' average achievement. This is comparable to NAEP's reports that White students' average achievement (156 points) is 29 points higher than Black students' average achievement.

Table 13. Grade 12 HLM Models Showing Relationship Between White–Black Achievement Gap in the NAEP 2010 Civics Assessment and Testing Conditions for Students in the Testing Conditions Study

				Number of Problems in Sessions			Testing Condition Factors		
Grade 12	**Single Level Regression Model**	**Model 1 Student Level Gap**	**Model 2 Student Level Gap With School Control**	**Model 3 Without Controls**	**Model 4 Student Level Gap With School and Testing Conditions Controls**	**Model 5 with Student, School, Testing Conditions, and Testing Conditions × School Controls**	**Model 6 Without Controls**	**Model 7 Student Level Gap With School and Testing Conditions Controls**	**Model 8 With Student, School, Testing Conditions, and Testing Conditions × School Controls**
Intercept Predicted Mean Civics Score	157.1 (0.9)	156.2 (0.9)	157.4 (0.9)	153.4 (1.3)	157.2 (1.2)	157.1 (1.3)	151.9 (0.9)	157.0 (0.9)	157.0 (0.9)
Student-Level Variable									
Black	-27.3*** (1.9)	-23.3*** (1.3)	-21.3*** (1.4)		-21.3*** (1.4)	-21.3*** (1.4)		-21.4*** (1.4)	-21.4*** (1.4)
Minority School									
More Than 50% of Students Black or Hispanic			-8.6*** (2.1)		-8.7*** (2.2)	-8.5** (4.1)		-7.3*** (2.1)	-7.6*** (2.3)
Number of Problems (Reference Category <2 Problems)									
2–5 Problems				-1.7 (1.9)	1.3 (1.7)	1.3 (1.8)			
6 or More Problems				-12.3** (5.7)	-8.8* (4.9)	-8.2 (6.0)			
Interaction Between Number of Problems and Minority School									
2–5 Problems × More Than 50% of Students Black or Hispanic						-0.1 (4.7)			
6 or More Problems × More Than 50% of Students Black or Hispanic						-2.9 (9.6)			
Testing Condition Factors (reverse coded)									
Orderly Environment							-5.4*** (1.4)	-4.3*** (1.3)	-4.4*** (1.7)
No Disruptions							-2.0** (0.8)	-1.5** (0.7)	-1.7** (0.8)
No Participation Problems							-1.5* (0.9)	-1.0 (0.8)	-1.1 (1.0)
Interaction Between Testing Condition Factors and Minority School									
Orderly Environment × More Than 50% of Students Black or Hispanic									0.6 (2.5)
No Disruptions × More Than 50% of Students Black or Hispanic									1.1 (1.8)
No Participation Problems × More Than 50% of Students Black or Hispanic									0.5 (1.8)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.

****p* < .01. ***p* < .05. **p* <.10

Standard errors for model parameters are shown in parenthesis.

Model 2 indicates the minority school effect at 12th grade is somewhat smaller (by about 2–3 points) than at grades 4 and 8, and the effect of the student being Black is larger (by about 5 points). Model 2 estimates average achievement would be 8.6 points lower for students assessed in minority schools than for those assessed in nonminority schools and that, after controlling for the session being in a minority school, Black students' average achievement would be 21.3 points lower than that of White students. The Model 2 estimate of the effect of the students being Black remains stable across Models 4 and 5, as well as across Models 7 and 8, which control for testing conditions.

The substantial negative effect (12.3 points) associated with being assessed in a session with 6 or more problems (Model 3) is reduced, but is still 8.8 points after controlling for students being Black and being tested in a minority school. But it no longer is significant, after controlling for the interaction of number of problems with sessions being in a minority school. The interaction effect is not significant.

The testing condition factors also show significant negative effects (Models 7 and 8). Although small, the effects of *Orderly environment* (4.4) and *No disruptions* (1.7) remain significant even after controlling for student being Black, testing session in a minority school, and the interaction of factor scores and the session being in a minority school.

Summary. Using the HLM results to examine whether less-than-optimal testing conditions may have affected NAEP's estimates of the White–Black gaps, it can be observed that the gaps remain relatively stable when controlled both for sessions being in minority schools and for sessions with poor testing conditions. The White–Black gap is reduced somewhat when controlling for the session being in a minority school, although less so with each higher grade (from 4 points at grade 4 to 2 points by grade 12). When controlling for testing conditions, the White–Black gap remains essentially unchanged. Mirroring the results found in the first set of HLM analyses, the models indicate that there is a testing-conditions effect beyond the school effect, and that improving testing conditions could improve achievement to some extent for both White students and Black students. At grades 4 and 8, for problem-prone sessions (6–10 problems), the testing-conditions effect is primarily taken up by the minority schools variable, but at grade 12 there is an additional detrimental effect of almost 9 points. When looking at the interaction between number of problems and the session being in a minority school, the effect is not significant except at grade 8, where it is quite large (20.9 points) for sessions with 6 or more problems. With regard to types of testing condition problems, being in sessions with participation problems has a slight negative effect at grades 4 and 8, whereas being in disruptive sessions and especially in sessions with a disorderly environment has an adverse effect at grade 12. There does not appear to be an interaction effect for the testing condition factors at any of the three grades.

White–Hispanic Gap Analysis

Grade 4 Analyses. Table 14 presents the 4th-grade results for the HLM analyses of the White–Hispanic gap, which resemble the results found for the White–Black gap. The simple regression predicts a White–Hispanic average civics achievement gap of

24.6 points at 4th grade (compared to NAEP's 27 points). After controlling for the difference between schools (30 percent of the variance), Model 1 estimates Hispanic students would score 18.4 points lower than White students. Model 2 predicts that students tested in schools with more than 50 percent Black or Hispanic students would score 13.3 points lower than those not tested in minority schools and that controlling for the session being in a minority school would result in Hispanic students scoring 13.6 points lower than White students.

Table 14. Grade 4 HLM Models Showing Relationship Between White–Hispanic Achievement Gap in the NAEP 2010 Civics Assessment and Testing Conditions for Students in the Testing Conditions Study

				Number of Problems in Sessions			Testing Condition Factors		
Grade 4	**Single Level Regression Model**	**Model 1 Student Level Gap**	**Model 2 Student Level Gap With School Control**	**Model 3 Without Controls**	**Model 4 Student Level Gap With School and Testing Conditions Controls**	**Model 5 With Student, School, Testing Conditions, and Testing Conditions × School Controls**	**Model 6 Without Controls**	**Model 7 Student Level Gap With School and Testing Conditions Controls**	**Model 8 With Student, School, Testing Conditions, and Testing Conditions × School Controls**
Intercept Predicted Mean Civics Score	169.1 (0.8)	167.9 (0.8)	169.7 (0.8)	164.6 (1.2)	170.4 (1.1)	170.0 (1.1)	163.5 (0.8)	169.6 (0.8)	169.6 (0.8)
Student-Level Variable									
Hispanic	-24.6*** (1.5)	-18.4*** (1.3)	-13.6*** (1.4)		-13.6*** (1.4)	-13.7*** (1.4)		-13.7*** (1.4)	-13.7*** (1.4)
Minority School									
More Than 50% of Students Black or Hispanic			-13.3*** (1.8)		13.2*** (1.7)	-11.6*** (2.4)		-13.0*** (1.7)	-13.5*** (1.8)
Number of Problems (Reference Category <2 Problems)									
2–5 Problems				-2.1 (1.7)	-1.4 (1.5)	-0.7 (1.7)			
6 or More Problems				-6.2* (5.2)	-3.5 (3.8)	-2.6 (4.3)			
Interaction Between Number of Problems and Minority School									
2–5 Problems × More Than 50% of Students Black or Hispanic						-3.2 (3.2)			
6 or More Problems × More Than 50% of Students Black or Hispanic						-3.2 (8.7)			
Testing Condition Factors (reverse coded)									
Orderly Environment							-0.5 (0.8)	-0.5 (0.6)	-0.5 (0.7)
No Disruptions							-1.3 (1.1)	-1.4 (0.9)	-1.2 (1.0)
No Participation Problems							-3.0*** (1.0)	-2.0** (0.9)	-1.5 (1.0)
Interaction Between Testing Condition Factors and Minority School									
Orderly Environment × More Than 50% of Students Black or Hispanic									0.5 (1.7)
No Disruptions × More Than 50% of Students Black or Hispanic									-0.6 (1.6)
No Participation Problems × More Than 50% of Students Black or Hispanic									-2.2 (2.0)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.

***p < .01. **p < .05. *p <.10

Standard errors for model parameters are shown in parenthesis.

The estimated effect of the student being Hispanic is stable having controlled for testing conditions in Models 4 and 5 and 7 and 8. The minority-school effect also is stable in the presence of the testing condition factors.

Looking at the 4th-grade testing condition effects, the effect of number of problems in the session no longer is significant when in the presence of the student being Hispanic and being tested in a minority school (Model 4), whereas the effect of the *No participation* problems factor remains significant, although it is reduced from -3.0 to -2.0 points (Model 7). In Model 8, the interaction effects between the factors and sessions held in minority schools are not significant.

Grade 8 Analysis. Table 15 presents the 8th-grade results for the White–Hispanic gap analyses. The simple regression predicts a 21.3 point gap, compared to NAEP's 2010 report of 23 points. Model 1, controlling for differences between schools, predicts that Hispanic students would score 16.3 points lower than White students, and Model 2, controlling for the session being held in a minority school, further reduces the effect of the student being Hispanic to 13.5 points.

Table 15. Grade 8 HLM Models Showing Relationship Between White–Hispanic Achievement Gap in the NAEP 2010 Civics and Testing Conditions for Students in the Testing Conditions Study

				Number of Problems in Sessions			Testing Condition Factors		
Grade 8	Single Level Regression Model	Model 1 Student Level Gap	Model 2 Student Level Gap With School Control	Model 3 Without Controls	Model 4 Student Level Gap With School and Testing Conditions Controls	Model 5 With Student, School, Testing Conditions, and Testing Conditions × School Controls	Model 6 Without Controls	Model 7 Student Level Gap With School and Testing Conditions Controls	Model 8 With Student, School, Testing Conditions, and Testing Conditions × School Controls
Intercept Predicted Mean Civics Score	163.1 (0.9)	161.9 (0.8)	163.8 (0.9)	160.1 (1.1)	164.5 (1.1)	164.4 (1.1)	157.6 (0.9)	163.5 (0.9)	163.6 (0.9)
Student-Level Variable									
Hispanic	-21.3*** (1.6)	-16.3*** (1.4)	-13.5*** (1.4)		-13.4*** (1.4)	-13.4*** (1.4)		-13.3*** (1.4)	-13.3*** (1.4)
Minority School									
More Than 50% of Students Black or Hispanic			-12.2*** (1.9)		-11.9*** (1.9)	-11.1*** (2.7)		-11.4*** (1.9)	-11.2*** (1.9)
Number of Problems (Reference Category <2 Problems)									
2–5 Problems				-4.8*** (1.8)	-1.6 (1.6)	-1.8 (1.8)			
6 or More Problems				-6.6 (5.2)	-3.4 (3.5)	1.5 (3.5)			
Interaction Between Number of Problems and Minority School									
2–5 Problems × More Than 50% of Students Black or Hispanic						0.4 (3.7)			
6 or More Problems × More Than 50% of Students Black or Hispanic						-19.8*** (5.9)			
Testing Condition Factors (reverse coded)									
Orderly Environment							-1.2 (0.8)	-1.1* (0.6)	-0.9 (0.7)
No Disruptions							-2.4** (0.9)	-1.2 (0.7)	-0.9 (0.8)
No Participation Problems							-3.7*** (0.9)	-2.5*** (0.9)	-1.9** 0.9
Interaction Between Testing Condition Factors and Minority School									
Orderly Environment × More Than 50% of Students Black or Hispanic									-0.8 (1.4)
No Disruptions × More Than 50% of Students Black or Hispanic									-0.8 (1.7)
No Participation Problems × More Than 50% of Students Black or Hispanic									-2.3 (2.4)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.

****p* < .01. ***p* < .05. **p* <.10

Standard errors for model parameters are shown in parenthesis.

The effect is about the same when also controlling for number of problems in the session (Model 4) as well as when controlling for the interaction of problematic sessions being held in minority schools (Model 5), and when replicating the analyses using the testing condition factor scores (Models 7 and 8). The effect of the session being conducted in a minority school (12.2 points) also holds relatively constant across the models, even in Model 5, which estimates a large, significant, negative effect of 19.8 points for the interaction between being tested in a session with 6 or more problems and being tested in a minority school.

Just like in the 8th-grade White-Black gap analysis, Model 3 shows an adverse effect on achievement associated with being tested in a session with 2–5 problems, although being tested in a session with 6 or more problems is not significant (probably due to the small sample size). Across the results for the testing condition factors (Models 6–8), the negative effect (3.7 points) of the *No participation problems* factor remains significant, although controlling for the student being Hispanic and for testing in a minority school reduces the effect from 3.7 to 2.5 points, and controlling for the interaction with testing in a minority school reduces the effect to 1.9 points. Unlike the result for sessions with 6 or more problems, there was no large interaction effect between the factor scores and the minority school indicator.

Grade 12 Analyses. Table 16 presents the 12th-grade results for the White–Hispanic gap in average civics achievement. Consistent with NAEP reports, the 12th-grade White–Hispanic score gap is estimated to be smaller than at grades 4 and 8, and about 10 points smaller than the 12th-grade White–Black score gap. The simple regression predicts Hispanic students would score 17.5 points lower than White students on average, similar to NAEP reports that Hispanic students' average 2010 civics achievement was 19 points lower than White students' average achievement.

Table 16. Grade 12 HLM Models Showing Relationship Between White–Hispanic Achievement Gap in the NAEP 2010 Civics Assessment and Testing Conditions for Students in the Testing Conditions Study

				Number of Problems in Sessions			Testing Condition Factors		
Grade 12	**Single Level Regression Model**	**Model 1 Student Level Gap**	**Model 2 Student Level Gap With School Control**	**Model 3 Without Controls**	**Model 4 Student Level Gap With School and Testing Conditions Controls**	**Model 5 With Student, School, Testing Conditions, and Testing Conditions × School Controls**	**Model 6 Without Controls**	**Model 7 Student Level Gap With School and Testing Conditions Controls**	**Model 8 With Student, School, Testing Conditions, and Testing Conditions × School Controls**
Intercept Predicted Mean Civics Score	157.1 (0.9)	156.4 (0.9)	157.3 (0.9)	154.4 (1.2)	157.2 (1.2)	157.2 (1.2)	153.0 (0.9)	157.0 (0.9)	157.0 (0.9)
Student-Level Variable									
Hispanic	-17.5*** (1.5)	-15.7*** (1.5)	-14.1*** (1.6)		-14.1*** (1.6)	-14.0*** (1.6)		-14.1*** (1.6)	-14.1*** (1.6)
Minority School									
More Than 50% of Students Black or Hispanic			-6.6*** (2.0)		-6.9*** (2.1)	-6.6* (3.8)		-5.6*** (2.0)	-4.7** (2.2)
Number of Problems (Reference Category <2 Problems)									
2–5 Problems				-2.0 (1.7)	0.9 (1.6)	1.1 (1.8)			
6 or More Problems				-8.0 (5.3)	-7.2 (5.1)	-9.0 (5.9)			
Interaction Between Number of Problems and Minority School									
2–5 Problems × More Than 50% of Students Black or Hispanic						-1.2 (4.3)			
6 or More Problems × More Than 50% of Students Black or Hispanic						11.2 (9.3)			
Testing Condition Factors (reverse coded)									
Orderly Environment							-4.3*** (1.3)	-3.6*** (1.3)	-4.4*** (1.6)
No Disruptions							-2.3*** (0.8)	-1.9*** (0.7)	-1.9** (0.8)
No Participation Problems							-0.9 (0.9)	-1.0 (0.9)	-0.9 (1.0)
Interaction Between Testing Condition Factors and Minority School									
Orderly Environment × More Than 50% of Students Black or Hispanic									3.2 (2.5)
No Disruptions × More Than 50% of Students Black or Hispanic									0.1 (1.6)
No Participation Problems × More Than 50% of Students Black or Hispanic									-1.3 (1.9)

Note: Hierarchical linear models with random intercepts. Testing Condition Factors have been centered around the grand mean.

****p* < .01. ***p* < .05. **p* <.10

Standard errors for model parameters are shown in parenthesis.

In Model 1, controlling for between-school differences, the predicted effect of the student being Hispanic (15.7 points) is only a little less than in the simple regression, and Model 2 predicts a smaller negative impact from being assessed in a minority school than the previous analyses do. The minority-school effect is significant, but smaller than in the previous gap analyses (6.6 points), and the effect of the student being Hispanic is only somewhat smaller in the presence of the minority school indicator (14.1 points).

Consistent with the previous analyses, at 12th grade the estimated effect of the student being Hispanic remained stable when controlling for both indicators of testing conditions—number of problems and the three testing condition factor scores. The minority school effect is about the same for number of problems. For the testing condition factors (Models 6–8), the minority school effect is somewhat reduced (1–2 points), and effects for two of the testing condition factors—*Orderly environment* (4.4 points) and *No disruptions* (1.9 points)—remain significant, although small, in the presence of the student being Hispanic and the testing session being in a minority school.

Summary. In general, the results of the HLM analyses of the White–Hispanic achievement gap parallel those for the White–Black gap. The White–Hispanic gap is reduced somewhat when controlling for sessions in minority schools, but less so at each higher grade. When controlling for testing conditions, the White–Hispanic gap remains essentially unchanged for both measures—number of problems and the testing condition factors. Looking at the results for number of problems in a testing session while controlling for minority school, there are no significant effects. However, similar to the results for the White–Black gap at the 8th grade, there is a large negative interaction (19.8 points) between the number of problems and sessions being in minority schools.

For the testing condition factors, the models indicate that there is a testing-conditions effect beyond the school effect, and that improving testing conditions could improve achievement to some extent for students across racial/ethnic groups. For the White–Hispanic gap analyses, the *No participation problems* factor is significant at the 4th and 8th grades, while the *Orderly environment* factor is significant at grades 8 and 12, and both the *Orderly environment* and *No disruptions* factors are significant at the 12th grade. The interactions between the three testing-condition factors and the sessions being in a minority school were not significant at any of the three grades.

Conclusions

The testing conditions study confirms that NAEP assessed most students in 2010 in conditions that were consistent with best practices. This addresses the first of our research questions. Also, students assessed in large sessions did not necessarily have lower achievement on NAEP's 2010 civics, U.S. history, and geography assessments. However, in general, achievement was lower when the number of problems in a session increased, particularly when the session had as many as 6–10 problems. Fortunately, essentially no sessions had more than 10 problems.

In addition to the cumulative impact of numerous problems in the same session, it is apparent that not all problems are equal, and that some are more serious than others. Results varied by grade, but in general students had lower achievement when there were problems in setting up, problems getting students to the session, certain types of session locations, or interruptions. Students also had lower achievement when they were assessed in situations that were cramped, uncomfortably hot or cold, or noisy, or that had visual distractions or numerous disruptions. Finally, unless students were in sessions where administrators agreed "a lot" that students were quiet and focused, achievement was lower. Achievement was particularly low for the 2–8 percent of the students assessed in sessions where students were not orderly or focused.

Three factors summarized the testing conditions data quite well: *Orderly environment* (adequate space with focused students), *No disruptions* (quiet without interruptions, distractions, and disruptions), and *No participation problems* (no problems with getting students to the sessions, refusals, or students leaving the sessions). At all three grades, sessions conducted outside of traditional classrooms were more prone to disruptions and participation problems.

Looking across the three grades, it is clear that as one moves to the higher grades, problematic testing conditions have a small but steadily increasing negative impact on NAEP achievement. At 4th grade, across the civics, U.S. history, and geography assessments, almost all students were assessed under conditions consistent with best assessment practices. For the most part, 4th-grade students were assessed in classrooms and in sessions with few problems. About 2 percent of 4th-grade students were assessed in sessions with 6–10 problems, however, and the overarching problem at 4th grade appears to be one of maintaining order. The only negative factor score at 4th grade was on the *Orderly environment* scale. Also, achievement was lower for the 11–14 percent assessed in sessions where the administrators agreed only "a little" rather than "a lot" that the 4th-grade students were orderly and quiet or that they were focused on the assessment. For the 3–4 percent assessed in sessions where the administrators disagreed that students were orderly and focused, achievement was low (-.26 to -.35 normits).

At 8th grade, 7 percent of the students were assessed in sessions with 6–10 problems and there were negative average factor scores on all three testing condition scales. Maintaining order was a problem, as with 4th grade. In addition, however, there were problems with disruptions during the sessions.

The administrators reported that 16 percent of the 8th-grade students were in sessions with interruptions. Looking at distractions and disruptions, 12 percent were in sessions that were noisy because of some school activity, 7 percent were in sessions with visual distractions, and 8 percent were in sessions with numerous school disruptions (e.g., the intercom)—all associated with lower achievement. It appears that, by the 8th grade, school-wide disruptions and interruptions have become routine in some schools. Finally, at 8th grade, participation became an issue. In particular, 19 percent of the students were in sessions where there were problems in getting students to the sessions.

At grade 12, there were fewer problems in maintaining order than at grades 4 and 8. However, there were increased problems with disruptions and participation. At 12th grade, 6 percent of the students were assessed in sessions with 6–10 problems, and there was greater negative impact on achievement than at grade 8. The average scores on the *No disruptions* and *No participation problems* factors also were lower than the scores on these factors at grade 8.

Looking across the three grades, the data indicate that minority and economically disadvantaged students are disproportionately likely to be tested in less-than-optimal conditions. Furthermore, disproportionate percentages of Black, Hispanic, and economically disadvantaged students attend schools that are more than 50 percent minority and schools that are economically disadvantaged. The HLM analyses of testing conditions, controlling for school socio-demographics and student indicators of disadvantage, indicated that the negative impact of problems in testing conditions is present over and above the challenges presented by disadvantaged schools. In particular, the *No participation problems* factor was significant at all three grades. This factor included problems in getting the students to the session on time, student refusals, and students leaving the session (primarily to go to the bathroom and get a drink of water). The *Orderly environment* and *No disruptions* factors also were significant at grades 8 and 12 (characterized by noise, unfocused students, disruptions, and interruptions).

The HLM analyses of the White–Black gaps and White–Hispanic gaps in average civics achievement indicated that the poorer testing conditions in minority schools probably have little impact on the achievement gaps at the student level. There is, however, a negative effect of being assessed in a minority school. Primarily, the gap analyses underscored the persistence of the effects of the testing condition factors across both the White–Black and White–Hispanic gap analyses, especially the *No participation problems* factor at grades 4 and 8 and the *Orderly environment* and *No disruptions* factors at grade 12.

Recommendations

Although the testing conditions study showed that NAEP 2010 testing conditions were generally in keeping with best assessment practices, there were some sessions with as many as 6–10 problems. Two percent of the 4th-graders and 6–7 percent of the 8th and 12th-graders were assessed in these multiproblem sessions, and these students had substantially lower achievement than students assessed in sessions without problems. According to analyses of the testing conditions most associated with lower achievement, the most serious problems included the following:

- Getting students to the session on time
- Students leaving the sessions (primarily for bathroom breaks or drinks of water)
- Interruptions and disruptions, including noisy rooms and visual distractions (often the PA system)
- Students not focused on the assessment and being disorderly

Of further concern, the testing-conditions problems were more pronounced in disadvantaged schools, including those serving minority students (more than 50 percent of students Black or Hispanic) and poor students (more than 50 percent of students eligible for free or reduced-price lunch). This could be anticipated, since minority students and those in poverty have lower-than-average NAEP scores. However, the testing conditions study results indicated that poor testing conditions have a negative effect on NAEP achievement that extends beyond the effects of poor schooling. Also, students in these sessions had very low achievement.

Operations

As soon as possible, NAEP should take steps to reduce the problems in testing sessions for all students, concentrating on disadvantaged schools.

More advance work. Student and teacher motivation could play a role in overcoming problems; therefore, it may be advisable for the NAEP program to increase efforts to explain the impact of problematic testing conditions on students' scores. This could involve preparing additional brochures or information sheets. It also could include more effort to meet with school personnel in advance of the testing sessions to highlight the importance of ensuring that students arrive at their sessions on time and that disruptions are kept to a minimum during the testing.

More support in conducting sessions. Potentially problematic schools could be identified in advance, and additional staff assigned to help maintain order in the sessions. Also, perhaps school staff members could be deputized to assist in getting students to the sessions, ensuring that students understood they were not supposed to leave the session (even to go to the bathroom), and perhaps even monitoring the PA system during the actual testing session.

Research Agenda

NAEP needs good information about testing conditions in schools on a routine basis. NAEP also needs further understanding about the extent of the impact of less-than-optimal testing conditions on student achievement. It appears that the effect is larger for students assessed in disadvantaged schools, which include disproportionate percentages of minority and poor students. There also may be differences among states in the magnitude of problems in testing sessions.

Prepare reports of testing conditions on a routine basis. The session debriefing form should be updated, taking into account the questions developed for the testing conditions study and the findings of the study. The information currently collected could be more useful. For example, if there was a problem with getting students to the session, did this involve most of the students or only one or two? Serious consideration also should be given to collecting data about the role that student and teacher motivation play in overcoming assessment administration problems. The data from the session debriefing form, including the written comments, should be analyzed and reported in conjunction with each assessment.

Collect testing conditions data as part of the state and district assessments. For more in-depth analysis, data based on the updated session debriefing form should be collected as part of a state-by-state assessment year. This would permit state-by-state comparisons of testing conditions to determine whether there is variation in testing conditions across states, and how that variation might relate to achievement. The in-depth analysis should also include data from the Trial Urban District Assessments (TUDA). Because problematic testing conditions are more likely to be found in disadvantaged schools, the data collected as part of the TUDA could be extremely valuable in studying the impact of testing conditions on the average achievement of minority and economically disadvantaged students.

Conduct an experiment (working with the TUDAs). If possible, an experiment should be conducted to collect data about the impact on average achievement of problems in testing sessions in disadvantaged schools. Poor, disruptive schooling results in lower achievement. Also, poor, disruptive conditions in testing sessions can lead to lower achievement, but it is impossible to fully separate the testing condition effects from schooling effects without an experimental study. It might be possible to work with one or two urban school districts to (1) determine how difficult it would be to improve testing conditions, since that would require considerable school cooperation, and (2) if the testing conditions were improved substantially compared to "normal," to determine the effect on average achievement.

References

National Center for Education Statistics. (2011). *The nation's report card: Civics 2010* (NCES 2011–466). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2010/2011466.pdf

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *Hierarchical linear and nonlinear modeling: HLM for Windows* (Version 6.00) [Computer Software]. Lincolnwood, IL: Scientific Software International.

Appendix A: 2010 Session Debriefing Form

Supervisor initials _____

SESSION DEBRIEFING FORM

COMPLETE THIS FORM FOR EACH SESSION - REGULAR, ACCOMMODATION, AND MAKEUP.

School ID #: ___________ Region#_________

Person Completing Form: _________________________ Supervisor: _______________________

Session Number: ____ (e.g.HG0401, CM0801)

Session Date: ________ What day of the week was the session held? □ Mon □ Tues □ Wed. □ Thurs □ Fri

Session Start Time: ________

Other NAEP Staff Assisting with Session: ____

Observers Present: □ Pearson □ HumRRO □NSC □ ETS □ Westat

School Observers: □ Teacher □ Principal □ Other, specify__________________________

This session was a: □ Regular Session □ Accommodation Session
□ Makeup Regular Session □ Makeup Accommodation Session

The session was held in a: □ Classroom □ Auditorium □ Lunchroom
□ Media Center/Library □ Other, specify_______________

SESSION SUMMARY (Be sure to provide as much detail as possible.)

ITEM	YES	NO	N/A	DETAILS
Were there any problems setting up for this session?				
Were there any problems getting students to this session?				
Were there any problems with the session timing?				
Were there any problems with the session materials (including the distribution and use of ancillary items)?				
Were there any student refusals?				
Were there any students who left the session?				
Were there any problems using the NAEP calculators?				
Were there any problems with accommodations given in this session?				
Were there any students still working when the timer rang?				
Were there any problems with the location?				
Were there any interruptions?				
Other, specify				

REACTION TO SESSION

AUDIENCE	ATTITUDE	COMMENTS/COMPLAINTS
Students	□ Positive □ Mixed/Indifferent □ Negative □ Can't say	
School Staff	□ Positive □ Mixed/Indifferent □ Negative □ Can't say	
Other Observers	□ Positive □ Mixed/Indifferent □ Negative □ Can't say	

LOCATION (Indicate how much you agree with the following statements by checking the appropriate box)

Item	Agree a lot	Agree a little	Disagree a little	Disagree a lot
The seating arrangement provided adequate space for students to work and not be distracted by each other				
There was ample space for me to move around and monitor individual students				
The lighting was adequate				
The temperature was comfortable				
The room was noisy because of a school activity (e.g., recess, band practice)				
There were visual distractions (e.g., student(s) leaving the session, activities occurring outside the window)				
There were numerous school disruptions (e.g., PA announcements, messages delivered)				
The students were orderly and quiet				
The students were focused on the assessment				

Number of Students (to be completed after the Administration Schedule is completed)

How many students were in this room? ______

Overall, how well did this session go?

- ❑ Very well
- ❑ Satisfactory
- ❑ Unsatisfactory

If “Unsatisfactory,” record comment:

____

____

____

____

____

____

____

Record any UNUSUAL circumstances in this session not previously mentioned:

____

____

____

____

____

____

____

____

Record any questions that students asked during the session. Be sure to include the subject and booklet number for questions about items.

Subject	**Booklet ID #**	**Student Question**

RETURN THIS COMPLETED FORM TO YOUR SUPERVISOR.

Appendix B: Classification of NAEP Assessment Administrators' Comments Provided on Session Debriefing Forms

Were There Any Other Session Locations?

	Grade 4		**Grade 8**		**Grade 12**	
Number of Comments	**43**		**86**		**178**	
	Number of Comments	**Percent of Comments**	**Number of Comments**	**Percent of Comments**	**Number of Comments**	**Percent of Comments**
Nontraditional Classroom (e.g., Art Room, Music Room)	11	9	21	12	61	34
Open Space/Large Room (e.g., Multipurpose Room, Hallway)	12	9	61	34	66	37
Miscellaneous	20	16	4	2	32	18

Were There Any Problems Setting Up for This Session?

	Grade 4		**Grade 8**		**Grade 12**	
Number of Comments	**81**		**72**		**80**	
	Number of Comments	**Percent of Comments**	**Number of Comments**	**Percent of Comments**	**Number of Comments**	**Percent of Comments**
Problem With Room Assignment or Access	44	59	16	22	24	30
Problem With Room Setup (e.g., Too Crowded, Needed to Reorganize Furniture)	15	20	30	42	24	30
Teacher/Student Delays or Students Arrived Too Early	7	9	15	21	22	28
Miscellaneous (e.g., Changes to Accommodations, Testing Materials Missing, Noise)	15	20	11	15	10	13

Were There Any Problems Getting Students to This Session?

	Grade 4		**Grade 8**		**Grade 12**	
Number of Comments	**142**		**185**		**272**	
	Number of Comments	Percent of Comments	Number of Comments	Percent of Comments	Number of Comments	Percent of Comments
Problem With Room Change/Confusion About Room	112	79	28	15	37	14
Students Missing or Refusing to Take the Test	1	1	8	4	46	17
Students Late/Not Dismissed From Previous Session	7	5	133	72	152	56
Miscellaneous (e.g., School Event Taking Place, Weather Delay)	22	15	16	9	37	14

Were There Any Students Who Left the Session?

	Grade 4		Grade 8		Grade 12	
Number of Comments	**379**		**317**		**274**	
	Number of Comments	**Percent of Comments**	**Number of Comments**	**Percent of Comments**	**Number of Comments**	**Percent of Comments**
Student Left for Bathroom or to Get a Drink	317	84	269	86	213	78
Student Ill or Upset, Left to Doctor or School Nurse	45	12	22	7	20	7
Left for Other Commitment (e.g., Extracurricular Activity, Work, Called to Principal's Office)	6	2	11	4	22	8
Miscellaneous (e.g., Student Name or ID Number Not Listed)	11	3	10	3	19	7

Were There Any Students Still Working When the Timer Rang?

	Grade 4		Grade 8		Grade 12	
Number of Comments	**355**		**260**		**212**	
	Number of Comments	**Percent of Comments**	**Number of Comments**	**Percent of Comments**	**Number of Comments**	**Percent of Comments**
1 or 2 Students Did Not Finish	142	40	115	44	106	50
More Students or Unspecified Number of Students Did Not Finish	129	36	98	38	60	28
Some Students Did Not Finish a Section	48	14	28	11	26	12
Miscellaneous (e.g., Extended Time Was Given, Students Refused to Stop Working)	36	10	19	7	20	9

Were There Any Problems With the Location?

	Grade 4		Grade 8		Grade 12	
Number of Comments	**57**		**92**		**84**	
	Number of Comments	**Percent of Comments**	**Number of Comments**	**Percent of Comments**	**Number of Comments**	**Percent of Comments**
Cramped, Crowded, or Poorly Arranged Room	24	42	26	29	23	27
Noisy Room	14	25	35	39	30	36
Problems With Heating, Cooling, or Lighting	9	16	18	20	17	20
Miscellaneous (e.g., Location Hard to Find, Need for Other Activities)	10	18	10	11	14	17

Were There Any Interruptions?

	Grade 4		Grade 8		Grade 12	
Number of Comments	**127**		**160**		**178**	
	Number of Comments	**Percent of Comments**	**Number of Comments**	**Percent of Comments**	**Number of Comments**	**Percent of Comments**
Interruptions by Tested Students	14	11	7	4	6	3
Interruptions by Persons External to the Testing Entering the Room	46	36	62	35	42	24
Loud Noises from Intercom, Fire Alarm, Phones, etc.	59	46	104	58	121	68
Miscellaneous (e.g., Heating/Cooling Problems, Students Had to Get Materials)	8	6	6	3	9	5

Record of Unsatisfactory or Unusual Circumstances

	Grade 4		Grade 8		Grade 12	
Number of Comments	**245**		**190**		**190**	
	Number of Comments	**Percent of Comments**	**Number of Comments**	**Percent of Comments**	**Number of Comments**	**Percent of Comments**
Student Misbehavior	30	12	38	20	27	14
Setup Problems or Delays (e.g., Weather Delays, Room Changes)	20	8	14	7	21	11
Time Away From Test (e.g., Interruptions, Bathroom Breaks)	41	17	27	14	25	13
Issues With Test Procedures/Test Materials (e.g., Finishing Early, Not Enough Materials)	51	21	55	29	39	21
Miscellaneous	104	42	50	26	73	38

Because results are rounded to the nearest whole number, some totals may appear inconsistent.

Appendix C: Number of Sessions and Students in the Testing Conditions Study, by Assessment Subject

Table C-1. NAEP 2010 Civics Assessment

	Number of Sessions	Number of Students
Grade 4	738	5,554
Grade 8	468	7,917
Grade 12	391	7,799

Table C-2. NAEP 2010 History Assessment

	Number of Sessions	Number of Students
Grade 4	578	5,570
Grade 8	571	10,261
Grade 12	519	9,637

Table C-3. NAEP 2010 Geography Assessment

	Number of Sessions	Number of Students
Grade 4	576	5,574
Grade 8	569	8,194
Grade 12	516	7,746

Appendix D: Percentage of Students and Average Achievement (in Normits) for Each Type of Testing Condition Problem, by Assessment Subject

Table D-1. Percentage of Students and Average Achievement (in Normits) in the Testing Conditions Study of the NAEP 2010 Civics Assessment by Characteristics of the Testing Session, as Reported on the Session Debriefing Form

		Session Size							
		<20		20–40		41–60		>60	
	Overall Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	.08	26	.10	66	.07	4	.00	3	.08
Grade 8	.08	20	.09	78	.06	#	~	2	~
Grade 12	.05	42	.01	58	.08	0	~	#	~

	Session Location									
	Classroom		Auditorium		Lunchroom		Library		Other	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	85	.10	1	~	7	-.07	4	-.05	4	.08
Grade 8	49	.07	3	.11	11	.07	26	.10	11	.06
Grade 12	54	.05	3	.13	6	.00	18	-.04	19	.14

	Session Day									
	Monday		Tuesday		Wednesday		Thursday		Friday	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	7	.30	31	.09	26	.02	26	.03	9	.16
Grade 8	7	.15	25	.01	30	.08	28	.06	10	.23
Grade 12	7	.15	30	.01	27	.09	28	-.01	8	.20

	Grade 4		Grade 8		Grade 12	
Original Debriefing Form Questions	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Problems setting up	6	.02	8	.02	7	-.04
Problems getting students there	11	-.03	17	-.02	34	-.02
Problems with timing	1	~	2	.03	4	-.14
Problems with materials	1	~	1	~	#	~
Student refusals	2	~	5	-.07	19	.15
Students left session	34	.05	31	.03	24	.05
Problems with NAEP calculators	1	~	2	.60	0	~
Problems with accommodations	2	~	1	~	0	~
Students still working	63	.05	54	.03	51	.03
Problems with location	5	.18	7	.07	6	-.05
Interruptions	10	-.02	14	.01	19	-.06
Other	4	.23	1	~	1	~

New Study Questions	Agree a Lot		Agree a Little		Disagree a Little		Disagree a Lot	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Adequate Space for Students to Work								
Grade 4	75	.07	18	.13	6	.02	1	~
Grade 8	79	.08	13	.08	5	.03	2	.12
Grade 12	89	.07	7	-.13	3	-.16	0	~
Ample Space to Monitor Students								
Grade 4	84	.07	11	.15	4	.02	#	~
Grade 8	84	.08	10	.06	5	.22	2	~
Grade 12	89	.06	7	-.05	3	-.05	1	~
Lighting Adequate								
Grade 4	97	.08	2	.01	#	~	0	~
Grade 8	98	.08	2	~	#	~	0	~
Grade 12	98	.05	1	~	#	~	0	~
Temperature Comfortable								
Grade 4	83	.08	11	.11	4	.04	1	~
Grade 8	78	.10	15	.01	4	.15	1	~
Grade 12	82	.05	10	.01	5	.08	2	-.02
Room Noisy Because School Activity								
Grade 4	2	.09	4	.20	4	.04	89	.07
Grade 8	3	-.16	9	-.03	6	.06	81	.10
Grade 12	2	~	9	-.13	6	-.05	83	.08
Visual Distractions								
Grade 4	2	~	3	-.04	5	.13	89	.07
Grade 8	2	~	4	-.01	6	.06	88	.09
Grade 12	2	~	3	-.13	4	.09	90	.06
Numerous School Disruptions								
Grade 4	2	~	2	~	4	-.12	92	.09
Grade 8	2	~	5	-.06	7	-.03	85	.11
Grade 12	1	~	10	-.05	8	-.08	80	.08
Students Orderly and Quiet								
Grade 4	78	.11	15	.00	3	-.22	2	~
Grade 8	81	.12	11	-.05	6	-.07	1	~
Grade 12	91	.07	5	-.04	1	~	2	~
Students Focused on Assessment								
Grade 4	79	.13	16	-.10	3	-.15	1	~
Grade 8	80	.13	13	-.12	4	-.10	1	~
Grade 12	89	.08	9	-.11	1	~	1	~

How well did the session go?	Very Well		Satisfactory		Unsatisfactory	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	83	.13	14	-.17	1	~
Grade 8	82	.12	14	-.14	1	~
Grade 12	86	.08	12	-.03	1	~

Debriefing form responses missing for approximately 1% of the sessions at each grade.

~ Indicates insufficient data to report achievement.

# Rounds to zero.

Table D.2. Percentage of Students and Average Achievement (in Normits) in the Testing Conditions Study of the NAEP 2010 U.S. History Assessment by Characteristics of the Testing Session, as Reported on the Session Debriefing Form

		Session Size							
	Overall Average Achievement	<20		20–40		41–60		>60	
		Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	.07	39	.05	59	.06	2	.47	0	~
Grade 8	.09	9	.11	39	.02	50	.13	2	.33
Grade 12	.05	10	.06	38	-.05	44	.08	7	.34

	Session Location									
	Classroom		Auditorium		Lunchroom		Library		Other	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	84	.06	#	~	3	.21	6	-.02	7	.17
Grade 8	39	.05	4	.03	34	.13	13	.09	11	.15
Grade 12	25	.03	15	.14	24	.06	16	-.06	20	.08

	Session Day									
	Monday		Tuesday		Wednesday		Thursday		Friday	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	7	.33	33	.06	27	.00	26	.04	7	.19
Grade 8	7	.15	25	.02	32	.09	28	.10	9	.23
Grade 12	7	.06	28	.02	28	.10	29	-.01	9	.12

	Grade 4		Grade 8		Grade 12	
Original Debriefing Form Questions	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Problems setting up	9	-.01	9	.05	8	-.14
Problems getting students there	10	-.07	19	-.07	38	-.03
Problems with timing	2	-.10	4	.09	3	-.20
Problems with materials	2	~	2	~	1	~
Student refusals	2	.08	7	.13	27	.06
Students left session	32	.06	39	.12	45	.01
Problems with NAEP calculators	0	~	0	~	#	~
Problems with accommodations	#	~	1	~	1	~
Students still working	72	.07	57	.04	52	.05
Problems with location	4	-.15	10	-.06	12	.00
Interruptions	11	.12	17	.01	19	.05
Other	2	~	2	-.01	3	.07

	Agree a Lot		Agree a Little		Disagree a Little		Disagree a Lot	
New Study Questions	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**
Adequate Space for Students to Work								
Grade 4	80	.08	14	.02	4	-.05	2	~
Grade 8	79	.12	12	-.04	5	.20	3	-.15
Grade 12	80	.07	13	.00	4	-.07	2	-.24
Ample Space to Monitor Students								
Grade 4	87	.08	9	-.03	2	~	1	~
Grade 8	85	.10	9	.05	4	.14	1	~
Grade 12	84	.05	9	.15	4	-.15	2	-.11
Lighting Adequate								
Grade 4	97	.07	3	-.03	0	~	0	~
Grade 8	95	.09	3	.19	1	~	#	~
Grade 12	95	.05	3	.07	1	~	#	~
Temperature Comfortable								
Grade 4	82	.08	12	.04	5	.01	1	~
Grade 8	81	.11	12	.03	5	-.10	2	~
Grade 12	79	.09	13	-.09	5	-.30	2	~
Room Noisy Because School Activity								
Grade 4	1	~	3	.10	4	-.43	91	.08
Grade 8	2	-.07	10	.03	10	.00	76	.12
Grade 12	2	.05	8	-.08	7	.12	81	.06
Visual Distractions								
Grade 4	1	~	3	-.08	3	-.14	93	.07
Grade 8	2	.09	5	.05	6	-.02	85	.10
Grade 12	1	~	5	-.19	6	.01	88	.07
Numerous School Disruptions								
Grade 4	1	~	2	~	4	-.04	92	.07
Grade 8	2	0.1	6	.08	10	.10	80	.09
Grade 12	2	-.02	8	.04	8	.01	80	.06
Students Orderly and Quiet								
Grade 4	86	.11	9	-.17	3	-.27	1	~
Grade 8	80	.13	11	-.05	6	.04	2	-.34
Grade 12	89	.07	7	-.07	2	~	2	~
Students Focused on Assessment								
Grade 4	84	.11	12	-.11	2	-.48	1	~
Grade 8	77	.13	16	.04	4	-.28	2	~
Grade 12	84	.09	13	-.14	1	~	1	~

How well did the session go?	Very Well		Satisfactory		Unsatisfactory	
	Percent Students	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**
Grade 4	85	.09	11	-.10	#	~
Grade 8	83	.15	14	-.16	2	~
Grade 12	80	.09	15	-.12	1	~

Debriefing form responses missing for approximately 1% of the sessions at each grade.

~ Indicates insufficient data to report achievement.

# Rounds to zero.

Table D.3. Percentage of Students and Average Achievement (in Normits) in the Testing Conditions Study of the NAEP 2010 Geography Assessment by Characteristics of the Testing Session, as Reported on the Session Debriefing Form

		Session Size							
	Overall Average Achievement	<20		20–40		41–60		>60	
		Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	.06	39	.07	59	.04	2	.36	0	~
Grade 8	.09	9	.14	39	.04	50	.09	2	.44
Grade 12	.04	10	.00	38	-.02	44	.08	7	.21

	Session Location									
	Classroom		Auditorium		Lunchroom		Library		Other	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	85	.05	#	~	3	.54	6	-.13	7	.13
Grade 8	38	.07	4	.20	34	.13	13	-.01	11	.05
Grade 12	24	.01	15	.15	24	.00	16	.03	21	.04

	Session Day									
	Monday		Tuesday		Wednesday		Thursday		Friday	
	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Grade 4	7	.25	32	.03	27	.04	26	.02	7	.22
Grade 8	7	.09	24	-.01	31	.09	28	.09	9	.27
Grade 12	7	.21	29	.03	28	.08	27	-.04	9	.09

	Grade 4		Grade 8		Grade 12	
Original Debriefing Form Questions	Percent Students	Average Achievement	Percent Students	Average Achievement	Percent Students	Average Achievement
Problems setting up	9	.02	10	.04	9	-.21
Problems getting students there	10	-.07	19	-.11	38	-.04
Problems with timing	2	-.09	4	-.08	3	-.36
Problems with materials	2	~	2	~	2	~
Student refusals	2	.05	7	.14	27	.07
Students left session	31	.06	38	.08	44	.02
Problems with NAEP calculators	0	~	0	~	#	~
Problems with accommodations	#	~	1	~	1	~
Students still working	72	.06	57	.01	51	.01
Problems with location	4	-.05	10	-.09	12	-.01
Interruptions	10	.12	17	-.04	19	.00
Other	2	~	2	-.05	3	.16

	Agree a Lot		Agree a Little		Disagree a Little		Disagree a Lot	
New Study Questions	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**
Adequate Space for Students to Work								
Grade 4	80	.08	14	-.03	4	-.07	2	-.07
Grade 8	79	.11	12	.00	6	.11	2	-.32
Grade 12	80	.06	13	.01	4	-.05	2	-.33
Ample Space to Monitor Students								
Grade 4	87	.09	10	-.10	2	-.23	1	~
Grade 8	85	.09	9	.08	4	.08	1	~
Grade 12	84	.04	10	.21	4	-.25	2	-.20
Lighting Adequate								
Grade 4	97	.06	3	.04	0	~	0	~
Grade 8	95	.08	3	.17	1	~	#	~
Grade 12	95	.04	3	.10	1	~	#	~
Temperature Comfortable								
Grade 4	82	.08	12	.01	5	-.04	1	~
Grade 8	81	.10	11	.01	5	.00	2	~
Grade 12	79	.08	13	-.09	5	-.21	2	~
Room Noisy Because School Activity								
Grade 4	1	~	3	.13	4	-.46	91	.07
Grade 8	2	-.12	10	.05	10	-.07	76	.12
Grade 12	2	.03	8	-.21	8	.07	80	.06
Visual Distractions								
Grade 4	1	~	3	-.13	3	-.02	93	.06
Grade 8	2	-.06	5	.05	6	-.04	85	.10
Grade 12	1	~	5	-.28	6	.07	89	.06
Numerous School Disruptions								
Grade 4	#	~	2	~	4	-.14	93	.06
Grade 8	2	.13	6	.11	10	.08	80	.08
Grade 12	2	.12	8	-.09	8	-.06	80	.06
Students Orderly and Quiet								
Grade 4	86	.10	8	-.10	4	-.29	1	~
Grade 8	79	.14	11	-.14	6	-.02	3	-.30
Grade 12	89	.06	7	-.08	2	~	2	~
Students Focused on Assessment								
Grade 4	84	.11	12	-.17	2	-.45	1	~
Grade 8	77	.14	16	-.07	4	-.24	2	~
Grade 12	83	.08	13	-.13	1	~	1	~

How well did the session go?	Very Well		Satisfactory		Unsatisfactory	
	Percent Students	**Average Achievement**	**Percent Students**	**Average Achievement**	**Percent Students**	**Average Achievement**
Grade 4	85	.09	11	-.15	#	~
Grade 8	82	.14	14	-.20	2	~
Grade 12	81	.08	15	-.11	1	~

Debriefing form responses missing for approximately 1% of the sessions at each grade.

~ Indicates insufficient data to report achievement.

# Rounds to zero.