Using State Assessments to Assign Booklets to NAEP Students to Minimize Measurement Error: An Empirical Study in Four States

Donald H. McLaughlin
Beth A. Scarloss
Frances B. Stancavage
Charles D. Blankenship
American Institutes for Research

Commissioned by the NAEP Validity Studies (NVS) Panel
March 2005

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

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

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

David Grissmer
RAND

Larry Hedges
University of Chicago

Gerunda Hughes
Howard University

Robert Linn
University of Colorado

Donald M. McLaughlin
American Institutes for Research

Ina V.S. Mullis
Boston College

Jeffrey Nellhaus
Massachusetts State Department of Education

P. David Pearson
University of California, Berkeley

Lorrie Shepard
University of Colorado

David Thissen
University of North Carolina-Chapel Hill

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
1791 Arastradero Road
Palo Alto, CA 94304-1337
Phone: 650/ 493-3550
Fax: 650/ 858-0958

Table of Contents

Introduction

Because NAEP estimates of state-level achievement play an important role in the evaluation of strategies for improving the nation's educational system, it is important that these estimates have as small a standard error as possible. NAEP estimates for state-level achievement are based on measurement of the performance of a random sample of students, and the 40-minute measurement of each student's performance includes inherent random error. The standard error of state-level estimates can be reduced either by increasing the sample size, which is expensive, or by reducing the error in each student's measurement. The error of measurement can be reduced by increasing testing time, but that would also entail additional cost, as well as additional burden on the students selected to participate in NAEP.

The error of measurement varies from student to student, and that variation depends on the "fit" between the student and the test. For the highest achieving students, easy test items provide little information, and for the lowest achieving students, hard test items provide little information. NAEP test booklets each include two blocks of items, and these item blocks vary in difficulty. If booklets with at least one easy block could be targeted to include the lowest achieving students, then the error of measurement for the segment of the population they represent could be reduced at very little cost. Likewise, if booklets with at least one challenging block could be targeted to include the highest achieving students, their measurement error could be reduced. A savings of 10 percent in the measurement error would produce benefits equivalent to increasing the length of the test or the number of students tested by nearly 20 percent, and a simulation study sponsored by the NAEP Validity Studies Panel estimated that such a reduction should be possible (Linn, McLaughlin, Jiang, and Gallagher, 2004).

The standard error of NAEP achievement statistics tends to be larger, other things equal, for low achieving groups, so there is a special need for blocks of items that are better fitted to their achievement levels. NAEP has attempted to develop such item blocks since 2000, with limited success. Nevertheless, NAEP item blocks differ in difficulty, and the present study was undertaken to estimate the relation between block difficulty and average standard error of measurement, as well as the resulting reduction in the standard error if students were to be assigned an optimal item block based on the measurement of their achievement provided by state assessment scores. For this study, the records of participants in the 2003 NAEP reading and mathematics assessments in four states were matched to state assessment records, and the standard errors of lowest and highest quartile students, based on the state assessments, were compared for all of the existing NAEP item blocks.

Five research questions guided this study:

1. How different are the difficulties of blocks on existing NAEP reading and mathematics assessments?
2. Can state assessments identify potentially low achievers on NAEP?

3. Are standard errors affected by block difficulties; specifically, are the standard errors for predicted low-achieving students smaller when they are assigned a booklet with an easy block?
4. What is the impact of easier blocks on the standard errors for NAEP's demographic reporting groups?
5. What other factors, such as completion rate or block position may also influence standard errors for low-achieving students?

The question of feasibility of acquiring and using state assessment scores is addressed in a companion report (McLaughlin, Scarloss, Stancavage, & Blankenship, 2005), and is answered affirmatively.

1. *How different are the difficulties of blocks on existing NAEP reading and mathematics assessments?*

The two-stage testing strategy for reducing measurement error depends on the availability of alternative second stages (i.e., NAEP booklets) that vary in difficulty. For optimal information about a segment of the student population, the test items should be designed so that the average percent correct (referred to as "P+" for brevity) is about 2/3. Items that are optimal for the lowest achieving students might have values of P+ for the entire population of as much as 90 percent, while those that are optimal for the highest achieving students might have values of P+ not substantially greater than the guessing rate for the population as a whole. The first question is the extent of variation of the current NAEP block difficulties. Could substantially easier item blocks be generated, or is the current variation as great as can be obtained?

2. *Can state assessments identify potentially low achievers on NAEP?*

Administration of a "screener" test as a part of NAEP administration would dramatically increase the cost and complexity of NAEP, but pre-identification of potentially low-achieving students based on available test scores (i.e., state assessment scores) would not substantially increase the cost and complexity of NAEP testing sessions, compared to current procedures, which already pre-assign booklets to students. The important question for the effectiveness of pre-assignment based on state assessment scores (e.g., quartiles) is the extent to which these scores would identify different levels of achievers on NAEP. The question can be addressed by examining the NAEP performance of students in different quartiles of the state assessment score distribution in each state.

3. *Are standard errors affected by block difficulties; specifically, are the standard errors for predicted low-achieving students smaller when they are assigned a booklet with an easy block?*

NAEP records the "posterior standard deviation" for each student record, which is an estimate of the size of the error in measuring the student's achievement on NAEP. Although this is not strictly the same as classical standard error of measurement, it is practically equivalent, since it used in the same way as a standard error of measurement in computing standard errors of aggregate state-level summary statistics.

The question is to what extent blocks with higher values of P+ have smaller posterior standard deviations. Unfortunately, NAEP does not provide these statistics for individual item blocks, only for performance on the complete booklet of items, which consists of two separately timed item blocks. Therefore, to approximate the relationship between block difficulty and standard errors, we compare posterior standard deviations for booklets containing each individual block.

The posterior standard deviation is, of course, a function of both item blocks in the student's test booklet, and each item block is paired with every other item block in the NAEP system. This means that the posterior standard deviations associated with particular blocks are somewhat less varied than they would be if blocks were administered separately. However, the two-block balanced incomplete block ("BIB") spiral design is essential to the conduct of NAEP. Moreover, that design ensures that members of "alleged" low-achieving groups, even if assigned booklets with one easy block, also have opportunities to respond to one of the more difficult item blocks.

4. ***What is the impact of easier blocks on the standard errors for NAEP's demographic reporting groups?***

   NAEP does not ordinarily report achievement by quartiles, but rather by demographic reporting groups. Therefore, it is of more practical significance to know how much the standard errors might be reduced for demographic groups that are both relatively small and traditionally over-represented in the lowest quartile of achievement.

5. ***What other factors, such as completion rate or block position may also influence standard errors for low-achieving students?***

   It is a reasonable assumption that errors of measurement are greater for students who fail to complete a block of items. That effect may be completely confounded with the difficulties of the items in the block (i.e., with P+), or it may provide additional information for identifying item blocks that can yield smaller standard errors for students in the lowest quartile.

   Further, NAEP test booklets each consist of two separately timed blocks of items. It is possible that students in the lowest quartile do not respond reliably to an easy block if they first encounter a "de-motivating" difficult block. If that is the case, then it is important not only to assign a test booklet with an easy block to these students but to assign a booklet in which the easy block appears first.

Methods

NAEP Data

In NAEP, estimates of student achievement are first computed by subscale and then combined into a composite estimate. There are five subscales for mathematics, two subscales for grade 4 reading, and three subscales for grade 8 reading. NAEP items are presented in blocks, and each booklet contains two item blocks. For mathematics, each of the five subscales is represented in each of the blocks, as shown in table 1.

For reading, unlike for mathematics, each item block assesses a single subscale under the current assessment design. This limits the utility of identifying a single “easy” block for inclusion in reading booklets assigned to students who are optimally tested by easier items: the easy items would only be relevant to a single subscale, whereas the objective is improved precision on the composite reading achievement measure. Nevertheless, we carried out the analyses in this study for reading blocks, as well as mathematics blocks, in view of the possibility that the reading assessment might be redesigned at some point in the future.

Table 1. Numbers of items in each NAEP mathematics subscale in 2003, by block

Block:	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Grade 4										
Numbers and operations	7	6	3	9	10	7	8	8	8	9
Measurement	3	1	2	3	4	6	4	4	3	2
Geometry	4	4	5	4	1	2	1	1	3	1
Algebra	3	2	4	3	3	2	2	3	1	4
Data analysis/statistics	1	1	1	2	2	2	3	2	3	2
Grade 8										
Numbers and operations	4	5	4	9	5	5	6	4	5	4
Measurement	4	3	4	5	3	3	1	1	3	3
Geometry	3	6	4	5	3	3	3	3	3	3
Algebra	6	4	4	6	4	4	4	7	5	5
Data analysis/statistics	1	2	2	4	4	3	4	3	3	3

The main database for this study consisted of records from the 2003 NAEP grade 4 and grade 8 reading and mathematics assessments for four states. State assessment files were matched to the NAEP records to retrieve state reading and mathematics assessment scores for each student. The resultant database was used to address two distinct questions: the use of state assessments to assign booklets to NAEP students to minimize measurement error is considered in this report, while the use of state assessments to impute achievement of students absent from NAEP is considered in a companion report (McLaughlin et al., 2005). The companion report also addresses issues related to the feasibility of acquiring state assessment scores on a national scale and on a schedule commensurate with NAEP’s operational requirements.

Nearly all NAEP records were matched in the dataset prepared for this pair of studies. These included records for participating, absent, and excluded students. Only the records of participating students are relevant here, however. Across the four states and four assessments, there were 66,128 records of students who participated in the assessment and all but 72 were matched to state assessment records (an average match rate greater than 99.8 percent). However, 9,507 of the matched records were missing state assessment scores. The records used for this study were restricted to those with both NAEP and state assessment scores. These constituted 31,423 grade 4 records and 25,198 grade 8 records, spread roughly equally across the two subjects and four states.[1]

State Assessment Data

Because in practical use, time constraints are likely to dictate the need for state assessment scores from the prior school year, we attempted to acquire grade 3 and grade 7 state assessment scores. However, because not all states implemented reading and mathematics assessments in grades 3 and 7 during the years of this study, we used grade 4 and 8 scores where necessary. A brief summary of each state's test follows.

State test scores for State 1 report student performance, given as scaled scores, from the 2002 and 2003 State 1 Student Assessment (S1SA) in English/language arts and mathematics at fourth and seventh grades. The S1SA is a criterion-referenced test that uses three item types: selected response (multiple-choice); short constructed response; and extended constructed response. We also collected ITBS scores in reading and mathematics for the third grade; however, the S1SA scores were used in the analyses described in this report.

For State 2 mathematics and English/language arts test scores were obtained for grades 3 and 7 for 2001-2002 and 4 and 8 for 2002-2003. All test items used a multiple-choice format. Calculators were allowed for a preponderance of the mathematics items at both grade levels. Scores for grades 3 and 7 were used in the analyses described in this report.

For State 3 English/language arts test scores were obtained for grades 3 and 7 for 2001-2002 and grade 4 for 2002-2003. Mathematics scores were obtained for grades 4 and 8 in 2002-2003. The primary statistics in this report are based on the grades 4 and 7 for reading and 4 and 8 for mathematics.

State 4 reported scaled scores from the state's criterion-referenced Assessment of Student Competence in Reading/English/Language Arts and Mathematics from the 2003 administration at fourth and eighth grades. Scores were not available for grades 3 and 7 in 2002.

Analysis

The difficulty of each block was estimated by computing the weighted average of P+; that is, the average percentage of the maximum possible score achieved by students in the combined four-state sample. Dichotomous items were scored as one point for a correct answer, and partial credit items were scored in integers, from zero up to the maximum specified by

[1] A total of 28,428 records with paired NAEP and state assessment reading scores and 28,193 records with paired NAEP and state assessment mathematics scores were analyzed. The counts for the four states are, for grade 8, 4151, 7159, 7067, and 6821; and for grade 4, 6141, 7823, 8343, and 9116.

NAEP. Item scoring procedures provided by the Educational Testing Service to accompany the data were used for scoring.

Records were divided into four quartiles in each state, based on state assessment scores. The quartiles were unweighted. Analyses were carried out separately for each quartile, although results are only presented overall and for the lowest and highest quartiles.

To estimate the potential for reducing standard errors by introducing blocks with higher values of P+, a linear regression was estimated for the relation across blocks between P+ and the posterior standard deviation, which served as our measure of standard error. For these regressions, the posterior standard deviation values for each block were based on all of the booklets containing that block.

Posterior standard deviations are provided by NAEP for subscales, but not for the composite. It is important that any two-stage testing procedure minimize standard errors on all subscales, so statistics were computed for all subscales. However, for brevity of presentation, standard error estimates presented in this report are for composites. The posterior standard deviation for the composite for each record was computed from the subscale posterior standard deviations, using the observed correlations among the subscale mean plausible value deviations from their respective posterior means as estimates of the correlations among the subscale measurement errors. A second estimate of the standard error of measurement for each record was obtained as the standard deviation of the five plausible values. A third potential measure of standard error is the measurement error component of the standard error estimate for NAEP mean scores for students assigned to different booklets.

Analyses of the relations between block difficulty and errors of measurement were carried out using the first of these three standard error estimates, the posterior standard deviation recorded on the NAEP data files. The relations among these three standard error estimates were examined for one grade and subject, grade 8 mathematics, for both the four selected states and the entire 55 jurisdictions participating in the assessment.

Differences between "standard error" measures

As noted, the dependent variable used in the main analyses for this study was the "posterior standard deviation" provided by ETS on the plausible value file. The posterior standard deviation is the standard deviation used for generating the five plausible values for each student, and the variation of averages based on the five sets of plausible values is the basis for the measurement component of the reported standard errors of NAEP results.[2]

As a check on results in this paper, some analyses were recomputed using the other two standard error measures, the standard deviation of the plausible values and the measurement component of the standard error of the mean, as the dependent variable. Because the results were somewhat different, we calculated the correlation, across blocks, among these three

[2] The generation of plausible values involves an additional component of variance, related to the variance of the estimates of the conditioning (regression) coefficients, so the variance of the five plausible values for each student is slightly larger than the square of the posterior standard deviation. This is described in NAEP technical manuals.

measures and with P+, for the lowest quartile of students in each state in the 2003 NAEP grade 8 mathematics assessment. The averages of these correlations are in table 2, with the average correlations for the four states included in study shown below the diagonal, and the average correlations for all 55 jurisdictions shown above the diagonal. For example, the average of the correlations for the four study states, between the percent correct (P+) on an item block and the mean square of the posterior standard deviation for students in the lowest quartile in each state, whose booklets contain that block, was -0.41.

Table 2. Correlations of three standard error measures with P+, for students in the lowest quartile on NAEP in four states and in all jurisdictions

	Mean square of posterior standard deviation	Mean variance of plausible values	Squared measurement component of standard error of mean	Percent correct
Mean square of posterior standard deviation	1.00	0.84	0.14	-0.39
Mean variance of plausible values	0.82	1.00	0.14	-0.33
Squared measurement component of standard error of mean	0.06	0.07	1.00	-0.08
Percent correct	-0.41	-0.31	0.02	1.00

Source: The National Assessment of Educational Progress (NAEP) 2003 Grade 8 Mathematics Assessment
Note: Average correlations for 4 states included in study are below the diagonal, and average correlations for all 55 jurisdictions are above the diagonal. Entries are unweighted averages of weighted correlations for each state.

The use of the root mean square of the posterior standard deviation (labeled “PSD” in this report) appears to yield approximately the same results as if we had used the standard deviation of the plausible values. The average correlation, across blocks, of these two measures is 0.82 in the four states and 0.84 in all the states participating in the assessment; and the average correlation of the percent correct (P+) with these measures is similar (-0.31 and -0.41 in the 4 study states, and -0.33 and -0.39 in all states), although the correlation with the plausible value standard deviations is somewhat lower, probably due to both the introduction of random imputation variance and the inclusion of error variance in the conditioning coefficient estimates.

The correlations with the measurement error component of the estimated standard error of the mean, which is the standard deviation of the five means computed for the five different sets of plausible values, are very different. Unexpectedly, these standard error estimates are virtually uncorrelated with P+ or with the other standard error estimates.

The estimated standard error of a mean is a function of both the error of individual measurements and the sample size, but different sample sizes cannot be the explanation of the low correlations among these measures because the correlations are across blocks (n=10) and each block is administered to nearly exactly the same number of students. Also, all three of the standard error measures include effects of "conditioning," so the differences do not appear to be due to the inclusion of conditioning in some of the measures and not others. The reason for the low correlations may lie in the way in which the plausible values are constructed.

Results

1. How different are the difficulties of blocks on existing NAEP reading and mathematics assessments?

The first question concerns the range of difficulties of current NAEP item blocks in reading and mathematics. To address this question we computed P+ for each 2003 block, based on the four states in the study. The results are shown in tables 3 and 4. Table 3 displays the average percent correct (P+) for each block, based on all of the students in the study (NAEP students with state assessment scores). In this and all other tables in this report, the averages across four states are presented.

Table 3. Average percent correct in four states by block, by subject and grade

Block ID	Mathematics Grade 4	Block ID	Mathematics Grade 8	Block ID	Reading Grade 4	Block ID	Reading Grade 8
J	63.2	D	63.1	I	74.6	M	76.8
H	61.9	C	60.5	B	67.9	J	70.7
I	59.3	B	57.6	G	67.6	L	69.6
G	54.4	I	55.9	H	67.5	C	68.2
D	53.7	J	55.4	D	61.7	E	67.2
E	53.3	H	53.7	J	60.3	I	66.5
A	51.0	G	51.6	A	56.1	G	59.0
F	49.9	A	50.3	F	55.3	H	55.1
B	48.3	F	50.1	C	44.8	B	54.5
C	46.6	E	46.3	E	43.7	D	53.1
						F	52.4
						A	38.0

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments
Note: For Reading Grade 8, Block K is not included because it is a single block booklet.

Table 4 displays the P+ values based on students identified by the state assessment as being in the lowest quartile in achievement in each state.

Table 4. Average percent correct in lowest quartiles in four states by block, by subject and grade

Block ID	Mathematics Grade 4	Block ID	Mathematics Grade 8	Block ID	Reading Grade 4	Block ID	Reading Grade 8
J	45.2	D	44.6	I	58.1	M	60.3
H	41.0	C	41.6	G	50.8	J	56.4
I	40.0	B	37.7	H	48.1	L	51.1
D	37.0	I	37.4	B	45.1	E	46.7
E	35.6	J	35.2	J	40.5	I	44.7
F	34.4	F	32.6	D	36.6	C	44.7
G	33.0	A	31.2	F	36.6	G	40.8
A	31.2	G	31.1	A	33.6	B	37.5
C	28.0	H	31.0	C	27.3	D	36.5
B	25.0	E	26.8	E	26.9	F	35.7
						H	34.6
						A	18.2

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments
Note: For Reading Grade 8, Block K is not included because it is a single block booklet.

Although there is substantial variation in block difficulties for the population, there is also substantial room for generation of item blocks that are easier and harder than any in the 2003 assessment. The analyses in this study are limited to a range of P+ from 46 to 63 for mathematics and 38 to 77 for reading. Blocks in which the average score is 85 or 90 percent correct, in the overall population, should be scalable in NAEP. For the lowest quartile, none of the blocks in mathematics approach the level of optimal information. All have P+ values less than 50 percent. For these students, NAEP is clearly a difficult test. For reading, on the other hand, there appear to be one block at grade 4 and two at grade 8 that might be sufficiently easy for students in the lowest quartile.

2. *Can state assessments identify potentially low achievers on NAEP?*

The second question concerns the ability of state assessment scores to identify student populations that will achieve low scores on NAEP. To address this, we examined the NAEP mean scores for students in each quartile on each state's assessment, shown in table 5, and compared these means to the actual means of the quartiles as determined by NAEP, shown in table 6.

In both subjects and grades, students in the lowest quartile on state assessments clearly averaged lower performance on NAEP than other students, although not as low as the lowest quartile of the NAEP students.[3] It appears that state assessments can effectively identify

[3] Due to the substantial measurement error in NAEP at the individual student level, the variations in achievement between quartiles is overstated when the quartiles are defined by the same measure whose mean is shown.

students who will score low on NAEP. Roughly 60 percent (for reading) and 65 percent (for mathematics) of students in the lowest quartile on their state's assessment also had mean composite NAEP plausible values in the lowest quartile. Most of the remaining students in the lowest quartile on their state assessment were in the lower middle quartile on NAEP.

Table 5. Mean NAEP composite plausible value, by state assessment quartile, subject, and grade

State Assessment Quartile	Mathematics Grade 4	Mathematics Grade 8	Reading Grade 4	Reading Grade 8
Lowest Quartile	209.6	239.4	187.4	231.1
Lower Middle Quartile	228.5	265.6	211.7	255.5
Upper Middle Quartile	244.3	286.8	229.4	272.9
Highest Quartile	263.8	314.9	248.3	289.3

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Table 6. Mean NAEP composite plausible value, by NAEP quartile, subject, and grade

NAEP Quartile	Mathematics Grade 4	Mathematics Grade 8	Reading Grade 4	Reading Grade 8
Lowest Quartile	202.4	231.1	175.3	219.9
Lower Middle Quartile	227.7	266.0	209.0	253.7
Upper Middle Quartile	246.1	289.4	231.2	274.7
Highest Quartile	270.1	320.9	259.3	300.4

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

3. Are standard errors affected by block difficulties; specifically, are the standard errors for predicted low-achieving students smaller when they are assigned a booklet with an easy block?

Given that there is variation in block difficulties in NAEP and that state assessments can identify low scorers on NAEP, we consider the third question: the extent to which assignment of booklets with particular blocks has an effect on standard errors (as measured by the posterior standard deviations provided by NAEP). The range of posterior standard deviations, for students with booklets containing particular blocks, is shown in table 7. These posterior standard deviations are reported in the "theta" scale, in which the standard deviation of scores in the population is assumed to be equal to one.[4] We see, therefore, that the standard errors are in the range of approximately a third of a population standard deviation, slightly higher for the lowest quartile than for the highest quartile and slightly higher for reading than for mathematics.

[4] For mathematics and for grade 4 reading, the standard deviations of the composite posterior means across all students ranged from 0.77 to 0.93 in the four study states, and for grade 8 reading, they ranged from 1.14 to 1.26.

For the lowest quartile, the minimum posterior standard deviation is less than 10 percent less than the maximum, except for grade 8 mathematics, where it is 12 percent less than the maximum. Therefore, assignment of the lowest quartile students on the state assessment to booklets containing particular blocks cannot have a substantial effect with the current item blocks. New item blocks designed to provide accurate measures for low achieving populations are needed.

Table 7. Block minimum and maximum posterior standard deviations, for lowest and highest quartiles, by subject and grade

	Lowest quartile		Highest quartile	
	Minimum PSD	**Maximum PSD**	**Minimum PSD**	**Maximum PSD**
Mathematics Grade 4	0.267 (A)	0.289 (D)	0.238 (A)	0.255 (H)
Mathematics Grade 8	0.238 (D)	0.272 (E)	0.216 (F)	0.236 (B)
Reading Grade 4	0.345 (H)	0.377 (E)	0.303 (C)	0.359 (I)
Reading Grade 8	0.442 (I)	0.473 (J)	0.421 (A)	0.501 (M)

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: Block identifier is shown in parentheses. Values are computed based on all booklets containing a given item block.

To determine the extent to which designing blocks with easier items can improve the standard error, we examined the relationship between these variables. Twelve charts in the appendix show, for mathematics and reading, grades 4 and 8, scatter plots of the relation between P+ (percent correct) and PSD (posterior standard deviation) for the lowest quartile of students, the highest quartile of students, and all students combined. The results are summarized in table 8.

Table 8. Posterior standard deviations as a function of block percent correct, for lowest and highest quartiles, by subject and grade

	Lowest quartile			Highest quartile		
	Adjusted R^2	**PSD at P+=50**	**Rate of change, per P+ change of 1%**	**Adjusted R^2**	**PSD at P+=50**	**Rate of change, per P+ change of 1%**
Mathematics Grade 4	< 0	0.274	-0.00026	0.20	0.233	0.00056
Mathematics Grade 8	0.31	0.240	-0.00121	< 0	0.212	0.00058
Reading Grade 4	0.48	0.351	-0.00078	0.73	0.294	0.00152
Reading Grade 8	< 0	0.462	-0.00004	0.63	0.429	0.00174

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: Values are computed based on all booklets containing a given item block.

For students in the lowest quartile based on the state assessment, there is a tendency for students who received booklets with easier blocks to have smaller standard errors, while the reverse is true for students in the highest quartile. However, for the lowest quartile, the size of the effect, except for grade 4 reading and grade 8 mathematics, is very small. Based on extrapolation of the regression, for grade 8 mathematics, creation of a scalable block whose P+ is 85 for the whole population would reduce the standard error for the lowest quartile by 17 percent, which would be a valuable improvement. For grade 4 reading, the reduction would be 7 percent, which would be of questionable benefit, given the effort.

As an aside, for reading, these results also suggest that standard errors might be reduced for the highest quartile by assigning them booklets containing blocks with more difficult items. However, the block-subscale confounding of the reading assessments, mentioned earlier, must be considered in deciding whether to take action based on this finding.

The decision to focus on the lowest quartile in this study was largely arbitrary, although it was based in part on the idea that state assessment scores might be more easily available in terms of score categories (“achievement levels”), of which many states report four. The effect might be larger for a more extreme group, such as the lowest decile. In fact, however, the variation between blocks in P+ values for the lowest decile, shown in table 9, were found to be about the same as for the lowest quartile, although the average P+ values are lower.

Table 9. Average percent correct in four states by block, for the lowest decile, by subject and grade

Mathematics				Reading			
Grade 4		Grade 8		Grade 4		Grade 8	
Block ID	P+	Block ID	P+	Block ID	P+	Block ID	P+
J	41.6	D	38.9	I	52.3	M	53.1
I	35.8	I	34.4	G	44.5	J	49.4
H	34.9	C	34.3	H	40.0	L	42.3
D	33.3	B	29.9	B	35.7	E	38.9
E	31.3	J	29.5	J	35.0	C	38.7
F	31.0	F	28.8	F	32.4	I	37.6
G	26.7	G	26.7	D	29.8	G	32.2
A	25.5	A	25.6	A	27.9	B	32.0
C	23.9	H	24.9	C	24.5	F	31.7
B	21.4	E	22.4	E	24.1	D	30.9
						H	27.5
						A	12.1

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Standard errors are generally larger for the lowest decile than for the highest decile, as shown in table 10. The pattern of variation in posterior standard deviations between blocks, for the highest and lowest deciles shows again the variation of more than 10 percent for grade 8 mathematics, but for the lowest decile, the variation is also that large for grade 4 reading.

Table 10. Block minimum and maximum posterior standard deviations, for lowest and highest deciles, by subject and grade

	Lowest decile		Highest decile	
	Minimum PSD	**Maximum PSD**	**Minimum PSD**	**Maximum PSD**
Mathematics Grade 4	0.280 (H)	0.304 (D)	0.245 (A)	0.263 (H)
Mathematics Grade 8	0.250 (D)	0.288 (J)	0.221 (F)	0.244 (B)
Reading Grade 4	0.351 (I)	0.395 (E)	0.305 (C)	0.368 (I)
Reading Grade 8	0.455 (I)	0.496 (H)	0.432 (A)	0.514 (M)

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: Values are computed based on all booklets containing a given item block.

For grade 4 reading, as well as for grade 8 mathematics, easier blocks are associated with lower standard errors for the lowest decile, as shown in table 11. However, these effects are not much greater than for the lowest quartile, and if a two-stage strategy is implemented, it would have a larger impact through application to 25 percent of the students rather than 10 percent of the students.

Table 11. Posterior standard deviations as a function of block percent correct, for lowest and highest deciles, by subject and grade

	Lowest decile			Highest decile		
	Adjusted R^2	**PSD at P+=50**	**Rate of change, per P+ change of 1%**	**Adjusted R^2**	**PSD at P+=50**	**Rate of change, per P+ change of 1%**
Mathematics Grade 4	< 0	0.286	-0.00033	0.11	0.234	0.00064
Mathematics Grade 8	0.21	0.246	-0.00126	< 0	0.223	0.00031
Reading Grade 4	0.49	0.358	-0.00104	0.77	0.283	0.00194
Reading Grade 8	0.03	0.473	-0.00042	0.60	0.429	0.00190

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: Values are computed based on all booklets containing a given item block.

4. *What is the impact of easier blocks on the standard errors for NAEP's demographic reporting groups?*

To determine the effects for demographic groups, we carried out the same computations separately for Black, White, and Hispanic students. The results are shown in tables 12 (for mathematics) and 13 (for reading).

Table 12. Average percent correct in four states by block, for Blacks, Whites and Hispanics, for Mathematics by grade

Grade 4						Grade 8					
Black		White		Hispanic		Black		White		Hispanic	
Block		Block		Block		Block		Block		Block	
ID	P+	ID	P+	ID	P+	ID	P+	ID	P+	ID	P+
J	52.9	J	67.6	J	55.4	D	52.2	D	67.1	D	59.1
H	50.4	H	67.0	H	53.1	C	48.3	C	64.7	C	53.3
I	48.9	I	64.2	I	47.8	I	46.2	B	62.0	B	50.3
D	44.2	G	59.7	E	47.5	B	45.4	I	59.9	H	46.9
E	42.9	D	57.6	D	46.0	J	44.5	J	59.8	J	46.8
G	42.4	E	57.3	G	45.5	H	40.9	H	58.6	I	46.4
F	40.1	A	56.5	F	42.7	F	40.2	G	56.3	G	41.4
A	38.5	B	54.1	A	41.1	G	39.8	A	55.2	A	40.8
C	34.9	F	53.8	C	40.7	A	37.3	F	54.0	E	40.6
B	34.1	C	51.0	B	40.1	E	34.5	E	50.7	F	39.3

Source: The National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment

Table 13. Average percent correct in four states by block, for Blacks, Whites and Hispanics, for Reading by grade

Grade 4						Grade 8					
Black		White		Hispanic		Black		White		Hispanic	
Block		Block		Block		Block		Block		Block	
ID	P+	ID	P+	ID	P+	ID	P+	ID	P+	ID	P+
I	66.0	I	78.2	I	69.0	M	69.6	M	79.6	M	69.1
G	58.3	B	72.9	G	63.2	J	67.4	C	72.5	L	64.2
H	57.6	H	72.1	B	61.1	L	64.5	J	72.3	J	63.6
B	57.1	G	71.4	H	59.2	C	57.2	I	71.9	C	60.9
D	51.4	D	67.1	J	52.2	E	55.4	L	71.8	E	58.6
J	50.6	J	65.0	D	51.4	I	53.8	E	71.6	I	52.8
F	46.7	A	60.8	F	48.0	G	47.6	G	63.4	G	49.8
A	46.5	F	59.0	A	45.9	H	47.5	H	59.1	B	45.8
C	35.8	C	48.4	C	41.6	D	46.3	B	58.4	D	44.2
E	34.3	E	47.7	E	39.5	F	44.9	D	56.4	H	44.1
						B	44.4	F	55.8	F	42.6
						A	26.8	A	43.5	A	27.2

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading Assessment

The average P+ values for Black and Hispanic students are approximately midway between the average P+ values for all students and for the lowest quartile. Thus, providing easier booklets for the lowest quartile would provide easier booklets for some Black and Hispanic students, but the effects of easy blocks on standard errors for these groups would probably be smaller than the effects on the lowest quartile *per se*. The *variation* of average P+ values between blocks is about the same for Blacks and Hispanics as it is for the lowest quartile, slightly larger than for other students.

Variation between blocks in the posterior standard deviations for Black and Hispanic students follows a pattern similar that for the lowest quartile, as shown in table 14. The minimum PSD is more than 10 percent smaller than the maximum PSD only for grade 8 mathematics. If a particular block were to be identified for administration to the lowest quartile, it would be important that it be the same block for all reporting groups. As shown in table 14, the block with the minimal PSD is the same for all three racial/ethnic groups for mathematics, but not for reading, suggesting that use of the two-stage strategy might be more practical for mathematics than for reading, with the current block and booklet structure.

Table 14. Block minimum and maximum posterior standard deviations, for Blacks, Whites and Hispanics, by subject and grade

	Black		White		Hispanic	
	Minimum PSD	**Maximum PSD**	**Minimum PSD**	**Maximum PSD**	**Minimum PSD**	**Maximum PSD**
Mathematics Grade 4	0.259 (A)	0.278 (C)	0.239 (A)	0.253 (C)	0.253 (A)	0.279 (D)
Mathematics Grade 8	0.230 (D)	0.263 (J)	0.214 (D)	0.235 (J)	0.223 (D)	0.254 (J)
Reading Grade 4	0.341 (B)	0.364 (E)	0.316 (C)	0.346 (I)	0.334 (D)	0.351 (C)
Reading Grade 8	0.437 (I)	0.475 (J)	0.417 (A)	0.482 (M)	0.444 (A)	0.474 (J)

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: Values are computed based on all booklets containing a given item block.

The strength of the relationship between P+ and PSD is shown in tables 15 (for Black and White students) and 16 (for Hispanic and White students). The results are similar to those in table 8, for the lowest quartile, for Blacks and Hispanics, but not for Whites. That is, the two-stage strategy might reduce grade 8 mathematics standard errors for Black and Hispanic students but not for White students.

Table 15. Posterior standard deviations as a function of block percent correct, for Blacks and Whites, by subject and grade

	Blacks			Whites		
	Adjusted R^2	**PSD at P+=50**	**Rate of change, per P+ change of 1%**	**Adjusted R^2**	**PSD at P+=50**	**Rate of change, per P+ change of 1%**
Mathematics Grade 4	< 0	0.266	-0.00031	< 0	0.247	0.00003
Mathematics Grade 8	0.24	0.242	-0.00102	< 0	0.228	-0.00021
Reading Grade 4	0.24	0.353	-0.00042	0.52	0.321	0.00062
Reading Grade 8	< 0	0.457	0.00022	0.47	0.440	0.00112

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: Values are computed based on all booklets containing a given item block.

Table 16. Posterior standard deviations as a function of block percent correct, for Hispanics and Whites, by subject and grade

	Hispanics			Whites		
	Adjusted R^2	PSD at P+=50	Rate of change, per P+ change of 1%	Adjusted R^2	PSD at P+=50	Rate of change, per P+ change of 1%
Mathematics Grade 4	< 0	0.266	0.00001	< 0	0.247	0.00003
Mathematics Grade 8	0.42	0.240	-0.00092	< 0	0.228	-0.00021
Reading Grade 4	0.12	0.345	-0.00026	0.52	0.321	0.00062
Reading Grade 8	0.16	0.459	0.00042	0.47	0.440	0.00112

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: Values are computed based on all booklets containing a given item block.

5. *What other factors, such as completion rate or block position may also influence standard errors for low-achieving students?*

Completion rates. To investigate the impact of completion rates on standard errors, we computed the percent of students who completed each block, shown in tables 17 (for all students in four states) and 18 (for students in the lowest quartile on each state's assessment).

Table 17. Average percent completed in four states by block, by subject and grade

Block ID	Mathematics Grade 4	Block ID	Mathematics Grade 8	Block ID	Reading Grade 4	Block ID	Reading Grade 8
H	87.4	I	94.4	I	84.7	M	96.8
J	87.1	A	93.3	J	77.2	G	94.4
I	80	D	92.7	A	69.2	C	92.8
B	77.4	J	92	H	67.1	J	91.6
A	71.9	E	89.3	G	66.4	L	87.4
G	71.2	C	85.9	C	65.5	D	87.1
D	69.7	F	84.4	D	64.9	F	86
F	51.1	H	84.1	B	63.5	E	84.1
E	49.5	G	77.8	F	59.3	I	83.4
C	48.7	B	68.3	E	57.7	B	81.9
						A	78.2
						H	75.9

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments

Note: For Reading Grade 8, Block K is not included because it is a single block booklet.

Table 18. Average percent completed in lowest quartile in four states by block, by subject and grade

Block ID	Mathematics Grade 4	Block ID	Mathematics Grade 8	Block ID	Reading Grade 4	Block ID	Reading Grade 8
J	78.5	I	92.6	I	75.2	M	94.8
H	77	A	90.3	J	65.1	G	88.3
I	68.5	E	88.2	G	57.3	J	86.1
B	64.7	D	85.6	C	55.8	C	82.8
A	59.7	J	82.1	D	55.5	L	79.2
G	59.1	F	80.5	H	54.6	F	76.2
D	55.3	H	76.6	A	52.7	D	74.7
F	44.4	G	73.6	F	50.8	B	74.4
E	41	C	72.9	B	50.4	I	71.4
C	39.5	B	68.5	E	48.8	E	67.9
						A	65.4
						H	64.2

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments
Note: For Reading Grade 8, Block K is not included because it is a single block booklet.

There is a great deal of variation in the percentages of students who complete different blocks, both overall and in the lowest quartile. Comparing the patterns of completion with the patterns of P+ (shown in tables 3 and 4), it is clear that while there is some relation between completion percentage and P+ (blocks that more students complete tend to be easier blocks), the relation is not strong. In fact, for the grade 8 mathematics blocks, there is no relation between completion and P+.

Since these two potential predictors of variance in posterior standard deviations are not completely confounded with each other, it is reasonable to ask whether together they might better predict the PSDs than would the percent correct alone. However, comparing the results in table 19 with those in table 6, we see no improvement either in the R^2 values or in the size of the coefficients relating posterior standard deviations to P+. That is, adding completion percentages to the analysis does not improve the selection of blocks that produce smaller standard errors for the lowest quartile.

Table 19. Average posterior standard deviation as a function of percent correct and test completion, for the lowest quartile, by subject and grade

	Lowest quartile			
	Adjusted R^2	PSD at P+=50 and 100% completed	Rate of change, per P+ change of 1%	Rate of change, per completed change of 1%
Mathematics Grade 4	0.16	0.266	0.0002	-0.03700
Mathematics Grade 8	0.32	0.249	-0.0011	0.03991
Reading Grade 4	0.41	0.353	-0.0008	0.00620
Reading Grade 8	0	0.475	-0.0005	0.07317

Source: The National Assessment of Educational Progress (NAEP) 2003 Reading and Mathematics Assessments
Note: Values are computed based on all booklets containing a given item block.

Block position. To investigate the hypothesis that block difficulty is only important for blocks in the first position (because the performance of low performing students is degraded if they first confront a block that is substantially too difficult), we repeated our analyses comparing blocks only when they were presented in the first position.

These analyses were only carried out for the mathematics assessment item blocks. Variation in P+, shown in tables 20 (for all students) and 21 (for students in the lowest quartile in each state), is virtually the same for the blocks in position 1 as for the blocks in either position (see tables 3 and 4).

Table 20. Average percent correct in four states by block, by grade (analysis restricted to booklets in which the block appears in the first position)

Block ID	Mathematics Grade 4	Block ID	Mathematics Grade 8
J	64.0	D	63.2
H	62.4	C	62.0
I	59.2	B	58.3
G	54.9	J	56.7
D	53.9	I	56.0
E	53.6	H	55.1
A	50.7	A	52.3
F	50.1	G	52.1
B	48.3	F	50.6
C	47.9	E	46.3

Source: The National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment

Table 21. Average percent correct in lowest quartile in four states by block, by grade (analysis restricted to booklets in which the block appears in the first position)

Block ID	Mathematics Grade 4	Block ID	Mathematics Grade 8
J	46.5	D	44.9
H	42.3	C	43.0
I	40.9	B	38.1
D	36.9	I	37.2
E	36.2	J	36.4
F	34.9	F	33.4
G	33.9	G	32.2
A	32.0	A	31.9
C	28.5	H	31.8
B	25.3	E	28.2

Source: The National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment

Furthermore, the variation in posterior standard deviations between booklets with different blocks in position 1 is virtually the same as the variation in PSDs between booklets with those blocks appearing in either position, as can be seen by comparing the results in table 22 with those in table 7. Thus, there is no apparent advantage to presenting the easy block in position 1. In fact, these results are surprisingly similar, because the entries in table 6 are based on overlapping sets of booklets, while those in table 22 are not.[5]

Table 22. Block minimum posterior standard deviations, for lowest quartile, by grade (analysis restricted to booklets in which the block appears in the first position)

	Lowest quartile	
	Minimum PSD	Maximum PSD
Mathematics Grade 4	0.262 (H)	0.290 (B)
Mathematics Grade 8	0.240 (D)	0.273 (E)

Source: The National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment
Note: Values are computed based on all booklets containing a given item block in the first block position.

Finally, the relation between P+ and the posterior standard deviations, shown in table 23, was virtually the same when focusing on the block in position 1 as it was when focusing on booklets with the block in either position (see table 8).

[5] A fraction of the booklets would contain both the block associated with the minimum PSD and the block associated with the maximum PSD, and the computation of the average PSDs in table7 would include those booklets in the averages for both, thus artificially decreasing the variation between the smallest and largest average PSD.

Table 23. Block posterior standard deviations as a function of block 1 percent correct, for lowest and highest quartiles, by subject and grade (analysis restricted to booklets in which the block appears in the first position)

	Lowest quartile		
	Adjusted R^2	PSD at P+=50	Rate of change, per P+ change of 1%
Mathematics Grade 4	<0	0.274	-0.0003
Mathematics Grade 8	0.31	0.240	-0.0012

Source: The National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment
Note: Values are computed based on all booklets containing a given item block in the first block position.

Discussion

This study was undertaken based on the results of a simulation (Linn et al., 2004) that indicated the potential for improving the accuracy of NAEP measurement for low-achieving subpopulations by including blocks of items more closely targeted to their achievement levels (i.e., easier items). We proceeded to evaluate the two-stage approach by acquiring state assessment scores (after the fact) for the 2003 NAEP assessments in four states and comparing the standard errors of lowest quartile students, based on the state assessments, when they are administered different blocks of items on NAEP. The standard errors used in the analyses were the average posterior standard deviations provided on the NAEP plausible values files.[6]

The quartiles of students on state assessment tests were highly predictive of NAEP achievement: approximately two-thirds of the students in the lowest quartile on a state's assessment had NAEP plausible values in the lowest quartile of the NAEP sample in that state. It should be noted that the true strength of this association may have been greater: the observed strength of relation is limited by the large individual measurement error on NAEP.

The range of variation in block difficulties in the simulation was greater than the variation among existing NAEP item blocks. During the analysis of the actual results for 2003, it became apparent that to benefit from the two-stage approach, new NAEP item blocks would be required that were more different from current item blocks than the current blocks are from each other. For grade 8 mathematics, there was evidence that such a strategy would have improved the accuracy of estimation for low-achieving population groups, but for grade 4 mathematics and for reading assessments in both grades, there was no evidence in this study that assigning easier blocks to low-achieving students would have improved the accuracy of estimates for low-achieving population groups. Furthermore, the effectiveness of

[6] Three different measures of the standard error of NAEP scores are available, and, as a check on the analyses, correlations were computed between P+ and these three measures: the average posterior standard deviation provided on NAEP plausible value files, the average standard deviation of the five plausible values for a student, and the measurement component of the estimated standard error of the mean, computed using the standard NAEP methodology. The result was that the last of the three standard error measures was virtually uncorrelated with the first two and with P+. The reason for that finding was not immediately apparent.

the strategy was approximately the same, whether the target group was the lowest decile or the lowest quartile of students.

Other findings were that:

1. the effectiveness of the two-stage strategy for minimizing standard errors for traditionally low-achieving demographic “reporting groups” was approximately the same as for the artificial “lowest quartile” group;
2. the effectiveness of the strategy was not increased when percent completion was added to P+ (percent correct) as indicators of blocks that would have lower standard errors for low-achieving groups; and
3. the effectiveness of the strategy was not increased when the analysis was based only on the P+ of the block presented first on NAEP.

Certain logistical issues related to the current NAEP design also must be considered in evaluating potential benefits from two stage testing. First, based on the current assessment design, each NAEP mathematics block of items contains at least one item on each of the content subscales, but for reading, each block focuses on a single content subscale. Therefore, providing an easy reading block would only (directly) improve the accuracy of measurement for that subscale. To benefit from the two-stage strategy, either easy blocks would be needed for each reading subscale, or a block would be needed which included items from all subscales.

Secondly, implementation of two-stage testing of any kind would give rise to a situation in which NAEP items would no longer be randomly assigned with equal likelihood to all students in the NAEP sample.[7] While this would not interfere with the execution of the computer programs for scoring and reporting, there is a possibility that it would lead to a critical violation of some assumption on which that scoring and reporting is based. However, we have not identified any such essential violation of assumptions, and NAEP analytical methods are already designed to deal with non-equivalent groups of test-takers.

In summary, the goal of reducing standard errors for students in the extreme quantiles still appears worthy of merit. However, certain design challenges would have to be addressed in order to take advantage of two-stage testing. More importantly, the results of this empirical analysis suggest that new NAEP blocks with more extreme P+ values would be needed to reap the potential benefits of a two-stage testing strategy.

[7] Each student would have a non-zero likelihood of seeing each item, but some students would have a greater likelihood than others of seeing the items contained in an “easy” block.

References

Linn, B., McLaughlin, D., Jiang, T., and Gallagher, L. (2004). Assigning Adaptive NAEP Booklets Based on State Assessment Scores: A Simulation Study of the Impact on Standard Errors. Palo Alto, CA: American Institutes for Research.

McLaughlin, D. H., Scarloss, B., Stancavage, F.B., and Blankenship, C. (2005). Using State Assessments to Impute Achievement of Students Absent from NAEP: An Empirical Study in Four States

Appendix

Twelve graphs on the following pages display the relation between block percent correct and average posterior standard deviation, averaged over subscales and students. The first four graphs are for the lowest quartile of students in each state; the second four are for the highest quartile in each state; and the final four are for the entire set of students in each state.

The regression equation and R^2 accompanying each graph are simple, unweighted results computed by Microsoft Excel.

Figure A-1. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Lowest Quartile, Mathematics Grade 4
Figure A-2. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Lowest Quartile, Mathematics Grade 8
Figure A-3. Average Posterior Standard Deviation Across Two Subscales, as a Function of Block Percent Correct, for the Lowest Quartile, Reading Grade 4
Figure A-4. Average Posterior Standard Deviation Across Three Subscales, as a Function of Block Percent Correct, for the Lowest Quartile, Reading Grade 8
Figure A-5. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Highest Quartile, Mathematics Grade 4
Figure A-6. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Highest Quartile, Mathematics Grade 8
Figure A-7. Average Posterior Standard Deviation Across Two Subscales, as a Function of Block Percent Correct, for the Highest Quartile, Reading Grade 4
Figure A-8. Average Posterior Standard Deviation Across Three Subscales, as a Function of Block Percent Correct, for the Highest Quartile, Reading Grade 8
Figure A-9. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for All Quartiles Combined, Mathematics Grade 4
Figure A-10. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for All Quartiles Combined, Mathematics Grade 8
Figure A-11. Average Posterior Standard Deviation Across Two Subscales, as a Function of Block Percent Correct, for All Quartiles Combined, Reading Grade 4
Figure A-12. Average Posterior Standard Deviation Across Three Subscales, as a Function of Block Percent Correct, for All Quartiles Combined, Reading Grade 8

Figure A-1. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Lowest Quartile, Mathematics Grade 4

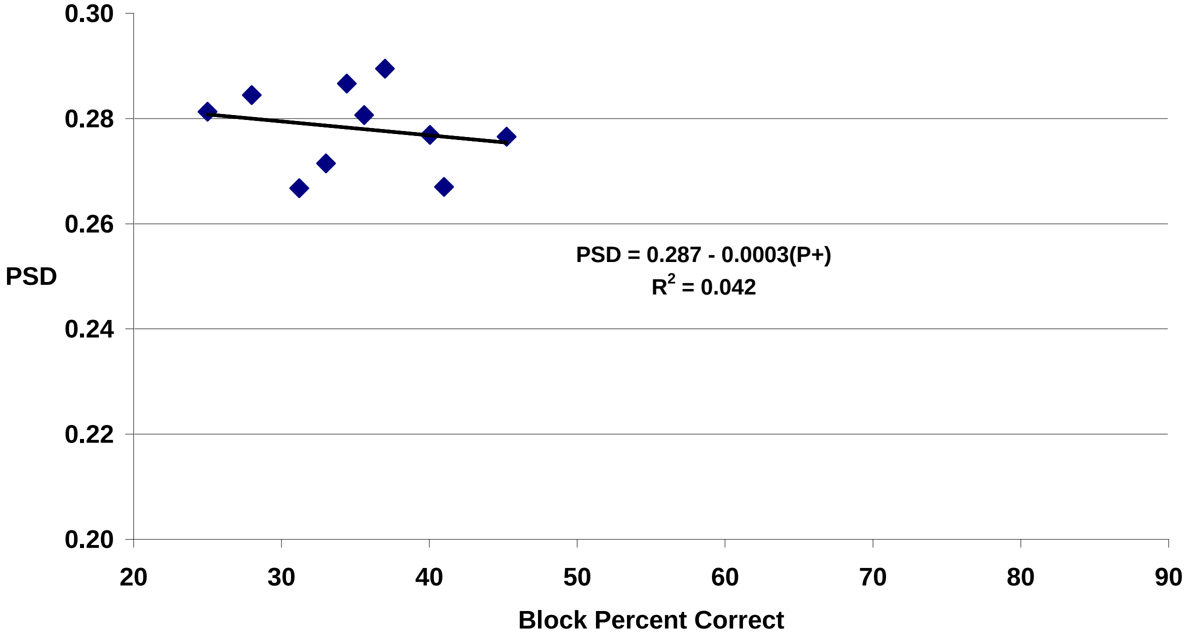


Figure A-2. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Lowest Quartile, Mathematics Grade 8

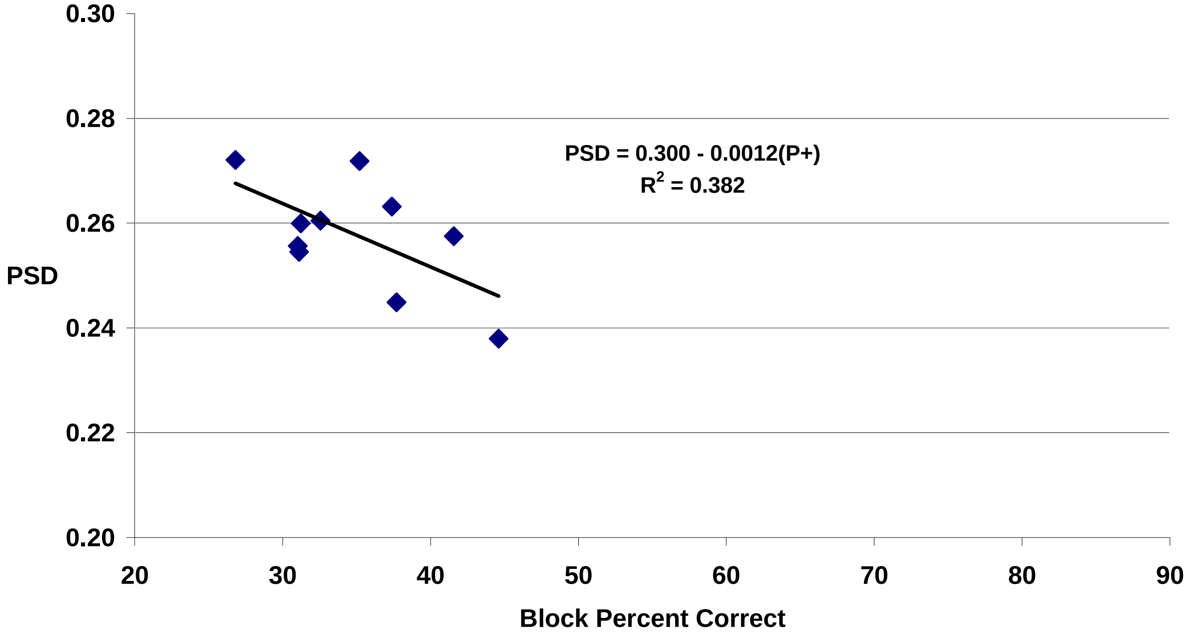


Figure A-3. Average Posterior Standard Deviation Across Two Subscales as a Function of Block Percent Correct, for the Lowest Quartile, Reading Grade 4

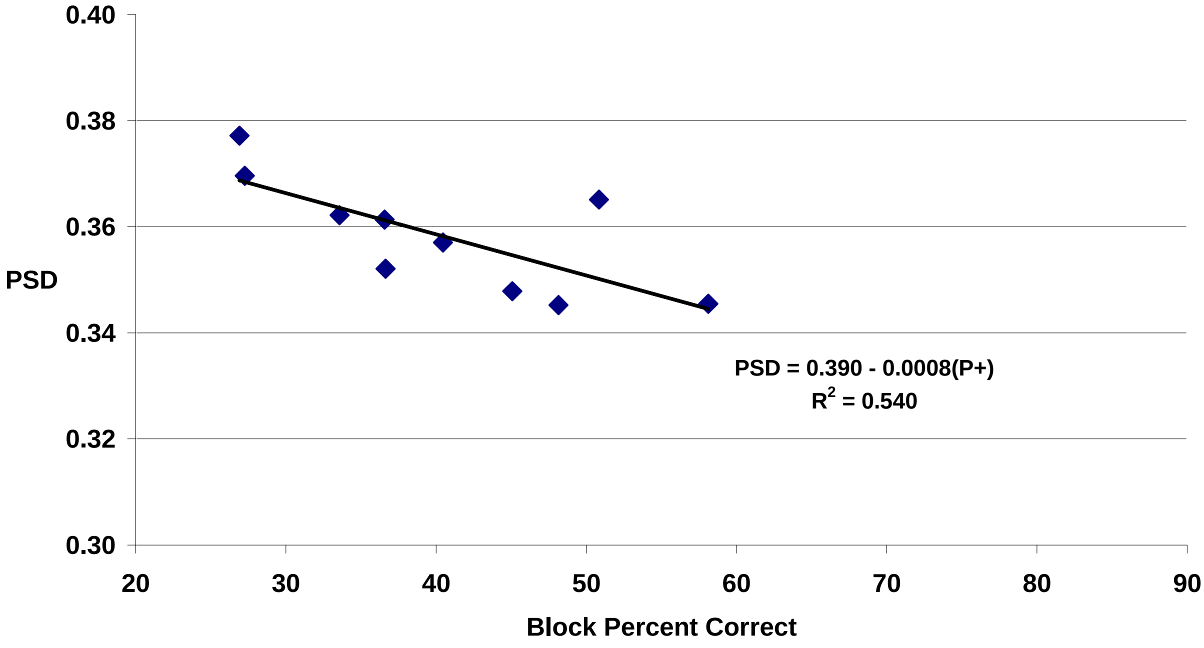


Figure A-4. Average Posterior Standard Deviation Across Three Subscales, as a Function of Block Percent Correct, for the Lowest Quartile, Reading Grade 8

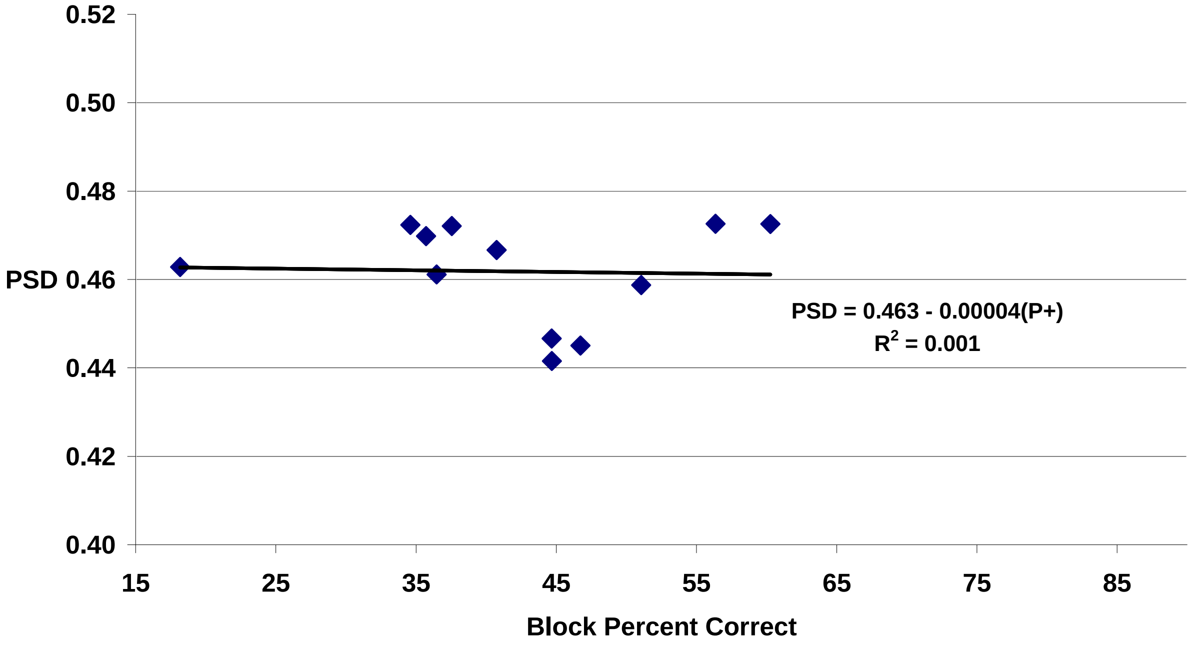


Figure A-5. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Highest Quartile, Mathematics Grade 4

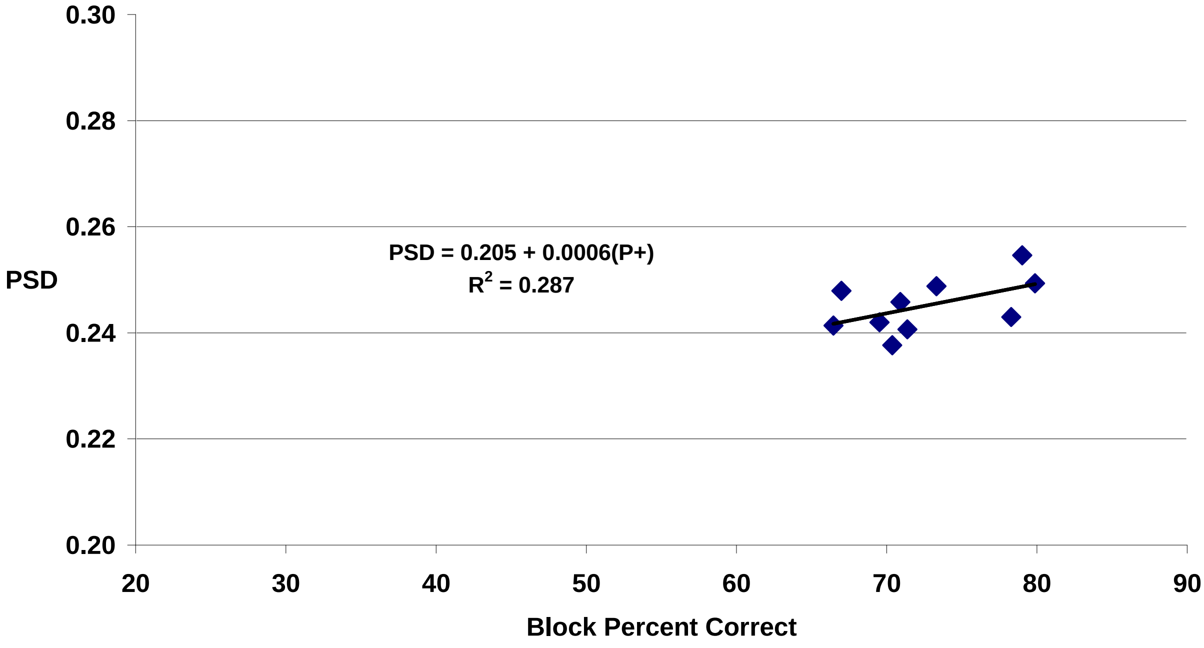


Figure A-6. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for the Highest Quartile, Mathematics Grade 8

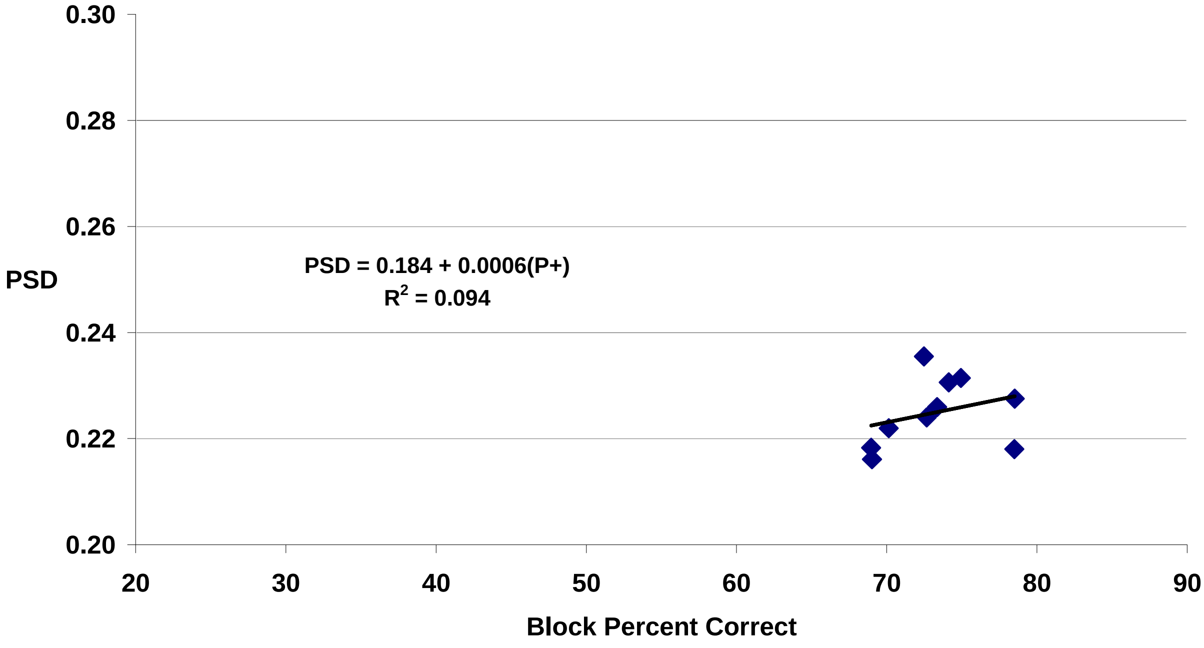


Figure A-7. Average Posterior Standard Deviation Across Two Subscales as a Function of Block Percent Correct, for the Highest Quartile, Reading Grade 4

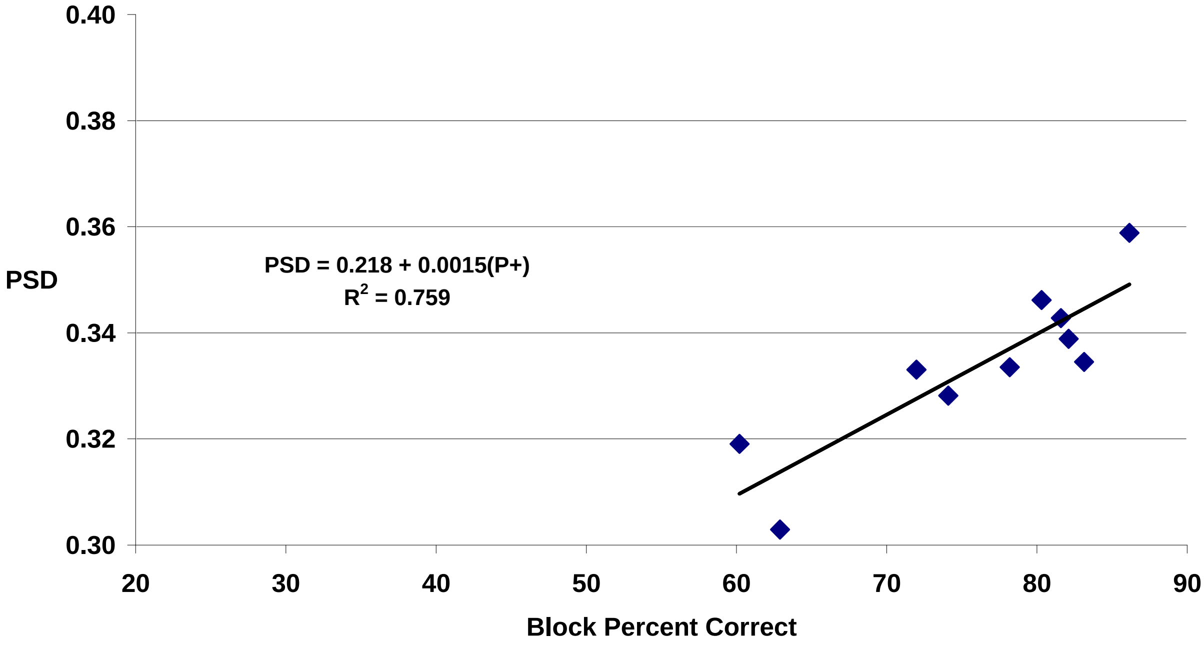


Figure A-8. Average Posterior Standard Deviation Across Three Subscales, as a Function of Block Percent Correct, for the Highest Quartile, Reading Grade 8

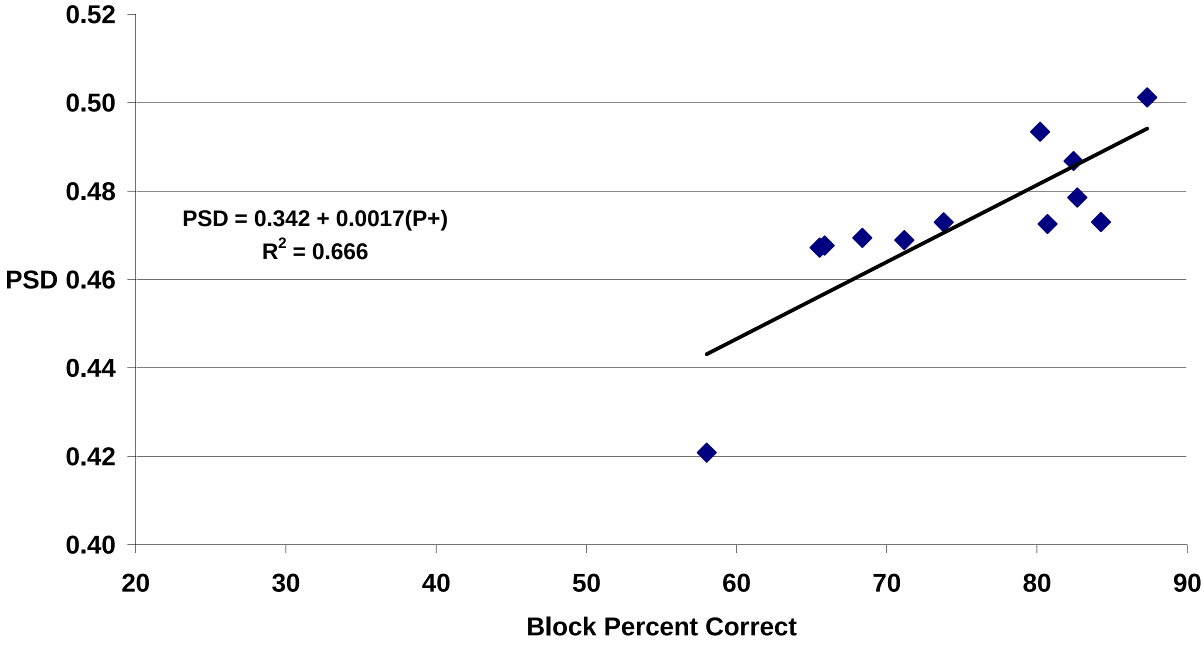


Figure A-9. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for All Quartiles Combined, Mathematics Grade 4

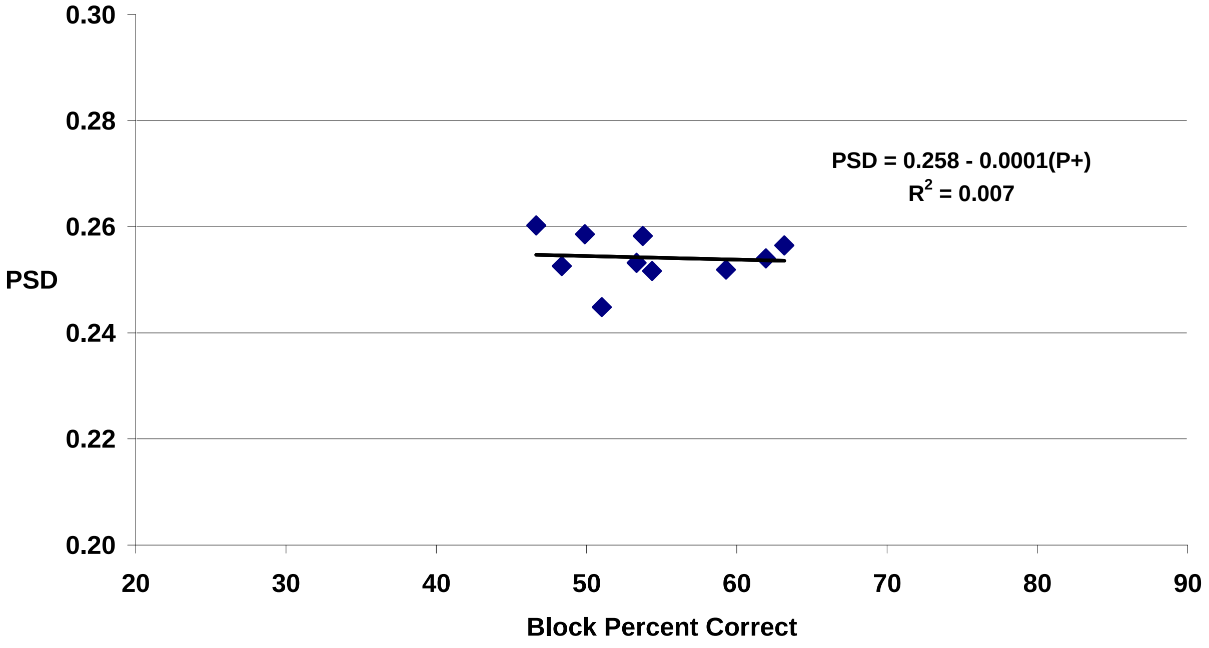


Figure A-10. Average Posterior Standard Deviation Across Five Subscales, as a Function of Block Percent Correct, for All Quartiles Combined, Mathematics Grade 8

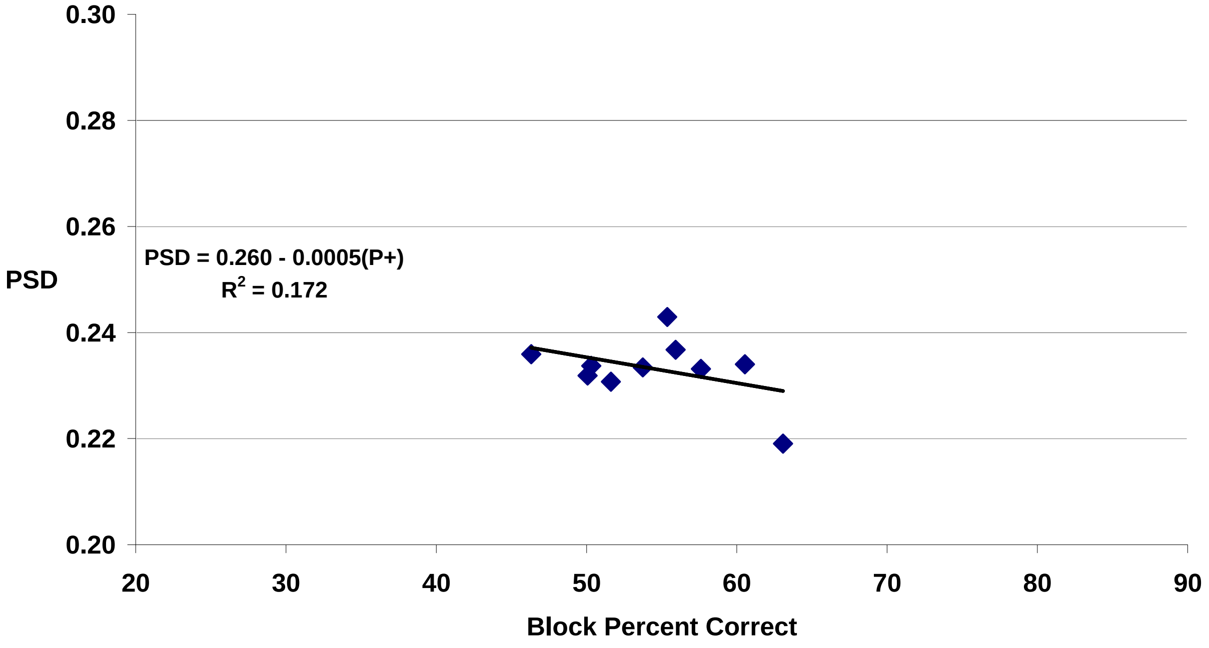


Figure A-11. Average Posterior Standard Deviation Across Two Subscales as a Function of Block Percent Correct, for All Quartiles Combined, Reading Grade 4

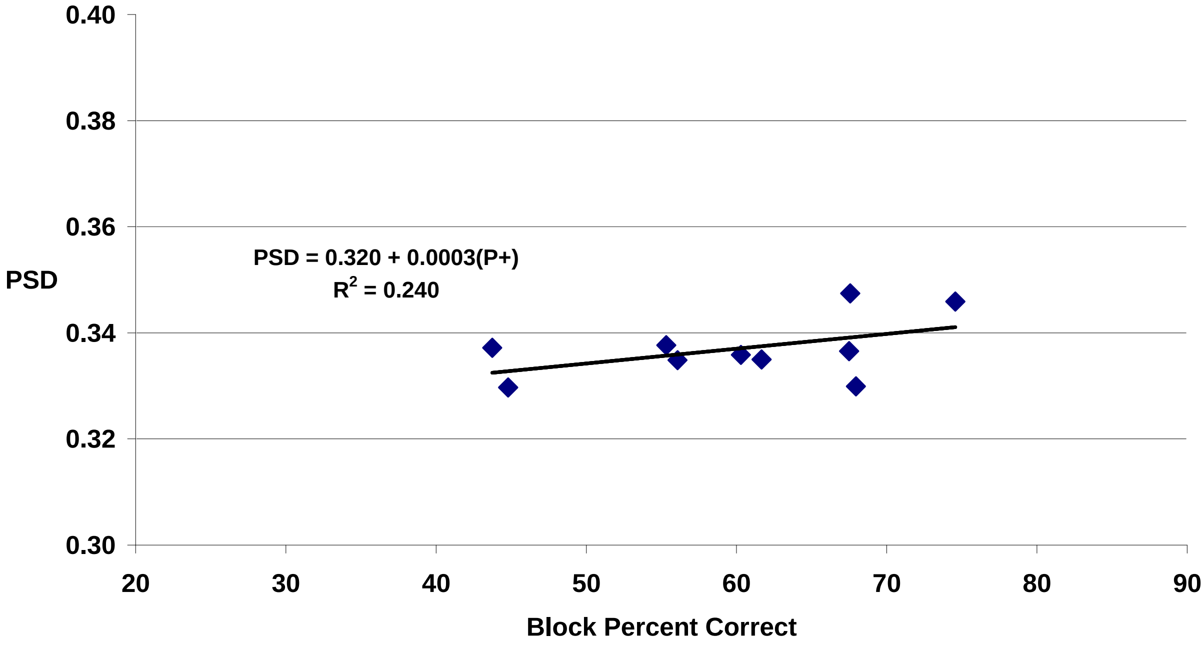
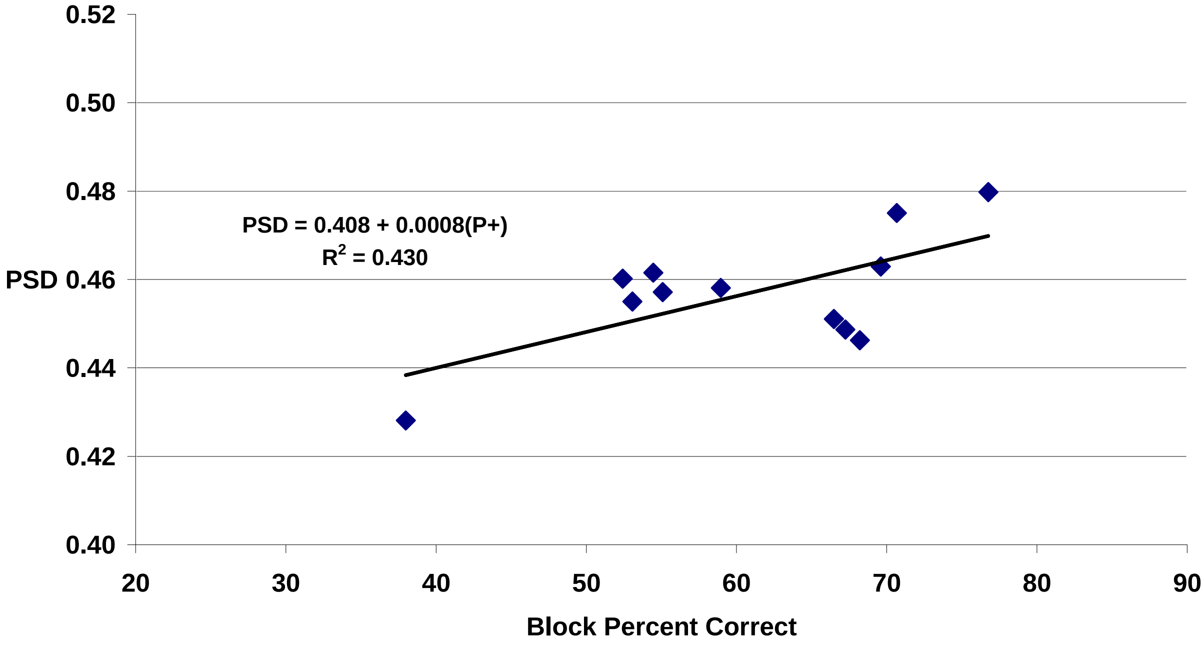


Figure A-12. Average Posterior Standard Deviation Across Three Subscales, as a Function of Block Percent Correct, for All Quartiles Combined, Reading Grade 8


PSD 0.46
PSD = 0.408 + 0.0008(P+)
R^2 = 0.430
Block Percent Correct