State of Delaware

2013–2014

Volume 4
Evidence of Reliability and Validity

American Institutes for Research

American Institutes for Research

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION AND OVERVIEW OF RELIABILITY AND VALIDITY EVIDENCE

The State of Delaware implemented a new online assessment for operational use during the 2010–2011 school year. This new test, referred to as the Delaware Comprehensive Assessment System (DCAS), replaced the paper-and-pencil test, referred to as the Delaware Student Testing Program (DSTP). As in previous school years (2010–2011, 2011–2012, and 2012–2013), students who were enrolled in various grades in public schools, including charter schools, were required to take the online assessment during the 2013–2014 school year. The paper-and-pencil version was available as an accommodation for students with special needs.

Both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic achievement from the DCAS test scores. This volume provides empirical evidence about the reliability and validity of the 2013–2014 DCAS, given its purported use.

Reliability refers to consistency in test scores. Reliability is directly tied to the standard errors of measurement (SEM)—the smaller the standard error, the higher the precision of the test scores. In item response theory (IRT), the standard errors of measurement vary from score to score—they are conditional. Because precision can be examined from various perspectives, this volume provides empirical data on precision in various ways, including marginal reliabilities, stratified alpha, mean standard errors by performance level, and mean standard errors by reporting subgroups.

Validity refers to the degree to which "evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Whether sufficient evidence has been presented to support the validity of a test is subject to professional judgment. Thus, organizing such evidence requires, first, an explicit statement regarding the intended uses of the test scores and, subsequently, evidence that the scores can be used to support these inferences.

The purpose of this volume is to provide empirical evidence to support the following:

- Content validity: Evidence that test forms are comparable across students and that each form aligns with the prioritized state content standards
- Internal structure validity: Evidence regarding the internal relationships among the subscale scores to support their use and the measurement model

2. PURPOSE OF DELAWARE'S STATE ASSESSMENT

The DCAS result serves as the primary indicator for the state's accountability system. As such, the test is a standards-based assessment designed to measure student achievement toward the state content standards. DCAS scores are indications of what students know and are able to do relative to the expectations by grade and subject area.

No test can measure the full breadth of the entire content standards. Rather, the standards and performance indicators measured by the DCAS are prioritized by the state to serve as indicators of an overall performance for students.

Unlike most state assessments, DCAS provides students with multiple opportunities through the school year to extend its potential to improve teaching and learning.

In 2013–2014, students participated in the DCAS test once in the fall and at least once in the spring. There are up to two test opportunities during the spring test window. Teachers could use the assessment data from the fall administration to adjust classroom instruction as needed and identify student growth between assessments.

Because pre-equating is used for the online test, test scores can be provided to students immediately upon completion of a test, given the existing item parameters. Teachers also have access to multiple score types for each student in an easy-to-access electronic data warehouse, including total scores, performance levels, and scores at the subcontent category level.

For linear fixed-form testing, test items are selected prior to the administration so that matching-to-test blueprints are ensured; while in computer-adaptive testing (CAT), test items are selected based on the blueprint for each student during the test. Wainer (2000) provides more extensive introductions to CAT. Since items are selected during the test administration, it is critical to show that each test in computer-adaptive testing conforms to the test blueprints. If the tests fail to match the test blueprint, the inferences regarding test scores could be reliable, but it would be difficult to establish validity evidence if the measured construct was not the same across individual test forms for students.

In DCAS, three types of student-level scores are generated: accountability score, instructional score, and strand score. Each of these scores has a specific intended use. Volume 1, Section 8.1, of the DCAS Annual Technical Report describes how each of these scores is computed.

The accountability score is computed based on the items measuring on-grade content only. This score is used to determine the performance level and proficiency status of each student by the Delaware Department of Education (DDOE) for accountability purposes. The instructional score is based on all items presented to the student—both on- and off-grade. This score is primarily used in monitoring the progress of groups of students over time.

The strand scores are also provided for each student as long as the strand has at least eight items per reporting content category. Scores are suppressed if fewer than eight items are presented to students within a strand. The purpose of these subscores is to indicate student strengths and weaknesses in different content areas of the test. These scores serve as useful feedback for teachers to tailor their instructions.

3. RELIABILITY

In computer-adaptive settings, items vary across students; thus, the marginal reliability (Thissen & Wainer, 2001) is reported for DCAS reading, mathematics, and science. *Marginal reliability* is a measure of the overall reliability of the test based on the average conditional standard errors, estimated at different points on the achievement scale, for all students. The marginal reliability coefficients are close to the coefficient alpha used in linear tests.

Within the IRT framework, measurement error varies across the range of ability as a result of the test information function (TIF). The TIF describes the amount of information provided by the test at each score point along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. The larger the measurement error, the less the information is being provided by the assessment at a specific level of ability.

According to the true score theory, the variance of observed scores consists of two orthogonal variance components, expressed as $var(x) = var(t) + var(e)$. True score theory also indicates that reliability is the ratio of true score variance to observed score variance, which can be expressed as

$$\text{reliability} = \frac{var(t)}{var(x)} = \frac{var(x) - var(e)}{var(x)}.$$

Then, the marginal reliability is defined as

$$\bar{\rho} = [\sigma^2 - \sum_{i=1}^{N} CSEM_i^2 / N) / \sigma^2 ,$$

where N is the number of students, $CSEM_i$ is the conditional SEM of the scaled score of student i, and σ^2 is the variance between students.

The fixed forms in social studies contain mixed-item types—constructed-response and multiple-choice. In these cases, it is appropriate to report the stratified Cronbach alpha coefficient computed as

$$\text{stratified } \alpha = 1 - \frac{\sum_{i=1}^{k} \sigma_i^2 (1 - \alpha_i)}{\sigma_x^2},$$

where α_i is the reliability of the ith strata, σ_i^2 is the variance between items in the ith strata, and σ_x^2 is the variance between all items on the test.

Table 1 presents the marginal reliability coefficients for each test by grade and test window based on the on-grade items for the accountability scores. Note that in reading and mathematics, for students taking both test opportunities in the spring 2014 window, the higher of two scores is used in these computations, as those scores count toward their accountability scores. Table 2 presents the stratified alpha for fixed-form-based tests (grade 2 in mathematics and reading; grades 4 and 7 in social studies) by grade and test form. The magnitude of reliability coefficients suggests that error variance accounts for about 8% to 17% of the total variance in scores. Larger conditional standard errors of measurement at the higher ends of the score distribution are the

primary source contributing to the larger error variance. Furthermore, selection of the higher of two scores from the spring 2014 window made this issue more apparent. As the item pool grows to include items that better target scores at the upper ends of the score distribution, the error variance will become smaller to yield a higher marginal reliability.

Table 1: Marginal Reliabilities of Accountability Scores

Grade	Mathematics		Reading		Science
	Fall Window	Spring Window	Fall Window	Spring Window	Spring Window
3	0.91	0.92	0.90	0.86	
4	0.90	0.91	0.88	0.83	
5	0.91	0.90	0.88	0.83	0.89
6	0.89	0.90	0.85	0.83	
7	0.89	0.90	0.87	0.83	
8	0.90	0.90	0.89	0.84	0.90
9	0.91	0.91	0.90	0.87	
10	0.89	0.89	0.87	0.84	0.90

Table 2: Stratified Alpha Coefficients of Accountability Scores

	Grade	Form(s)	Stratified Alpha
Mathematics	2	-	0.86
Reading	2	-	0.89
Social Studies	4	A	0.84
Social Studies	4	B	0.82
Social Studies	4	C	0.84
Social Studies	4	D	0.81
Social Studies	7	A	0.88
Social Studies	7	B	0.87
Social Studies	7	C	0.86
Social Studies	7	D	0.86

3.1 TEST INFORMATION CURVES AND STANDARD ERROR OF MEASUREMENT

In IRT, the TIF provides the information on how well the test measures student ability. With CAT, the TIF is unique for each individual test form that is tailored to the student's ability level. For CATs, instead of presenting the TIFs, it is more useful to examine the standard errors along the score distribution for each test and grade.

Theoretically, with an infinitely large item bank with a perfect match-to-ability to the population, standard error curves would be flat along the score range—an indication that all students are measured with the same precision. However, this is not practical because in the real world the item pools may not be perfectly balanced, especially in the early years of operation. Thus, the standard errors of measurement can be larger at either end of the distribution as more items are usually developed at the medium difficulty range.

It is useful to consider the conditional standard errors and their magnitude in the context of the item pool. If we knew how small the standard errors could be theoretically, we could evaluate whether the current item pool and algorithm select items appropriately to match a student's ability. Therefore, smaller standard errors of measurement would be provided for each student.

In IRT, the TIF provides a statistical indication of the information provided by a given set of items that make up the test. The TIF for a Rasch model is

$$TIF(\theta) = \sum_{j=1}^{N} I_j(\theta)$$

$$I_j(\theta) = \frac{\exp[\theta_i - b_j]}{(1+\exp[\theta_i - b_j])^2},$$

where j indicates item j ($j \in \{1,2,\ldots,N\}$), i indicates student i, θ_i is the ability of student i, and b_j is the difficulty parameter of item j.

Under the Rasch model, the standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

$$se(\theta_i) = \frac{1}{\sqrt{\sum_{j=1}^{N} I_j(\theta)}}.$$

To establish a theoretical framework for examining the efficiency of standard errors for CAT, assume we have an infinitely large item pool in order to always have a perfect match of the item parameter to a student's ability so that $\theta_i = b_j$ for all j. Then, the information provided by any given item is always

$$I_j(\theta) = \frac{\exp[0]}{(1+\exp[0])^2} = \frac{1}{4} = .25.$$

The standard error of theta would always depend on how many items selected, *N*, perfectly match the student's ability:

$$se(\theta_i) = \frac{1}{\sqrt{N * .25}}.$$

The DCAS adaptive tests are fixed with respect to test length. For example, if the test includes a total of 54 possible score points, the theoretical lower limit of the standard errors on the logit scale for the entire test (including both on-grade and off-grade items) is about

$$se(\theta_i) = \frac{1}{\sqrt{54 * .25}} = .272\,.$$

We use the formula above to derive the theoretical lower limit for each test and apply the linear transformation to place these results onto the DCAS reporting scale. Table 3 provides the results of the minimum achievable standard error on the DCAS reporting scale for the instructional score. Standard errors on the DCAS cannot be smaller than the values in this table, as these provide the theoretical lower bound.

Table 3: Theoretical Minimum Standard Errors by Grade and Subject

Grade	Reading			Mathematics			Science			Social Studies		
	# Items	Max Points	Lower Limit of SE	# Items	Max Points	Lower Limit of SE	# Items	Max Points	Lower Limit of SE	# Items	Max Points	Lower Limit of SE
2	30	30	25.9	30	32	22.6						
3	50	53	19.5	50	52	17.8						
4	50	53	19.5	50	52	17.8				46	48	11.4
5	50	53	19.5	50	53	17.6	50	54	12.8			
6	50	52	19.7	50	54	17.4						
7	50	53	19.5	50	54	17.4				46	48	15.7
8	50	52	19.7	50	54	17.4	50	54	13.4			
9	50	52	19.7	50	56	17.1						
10	50	52	19.7	50	54	17.4	50	54	14.3			

The figures below are plots of the conditional standard error of measurement for online test accountability scores obtained in spring 2014. The vertical line denotes the cut scores for the Below, Meets Standard, and Advanced cuts on each test. Note that only accountability scores are available in science and social studies as well as grade 2 mathematics and reading. Standard error plots for all test windows can be found in Volume 1, Appendix E.

Figure 1:
Conditional Standard Errors of Measurement (Mathematics), DCAS Spring Window

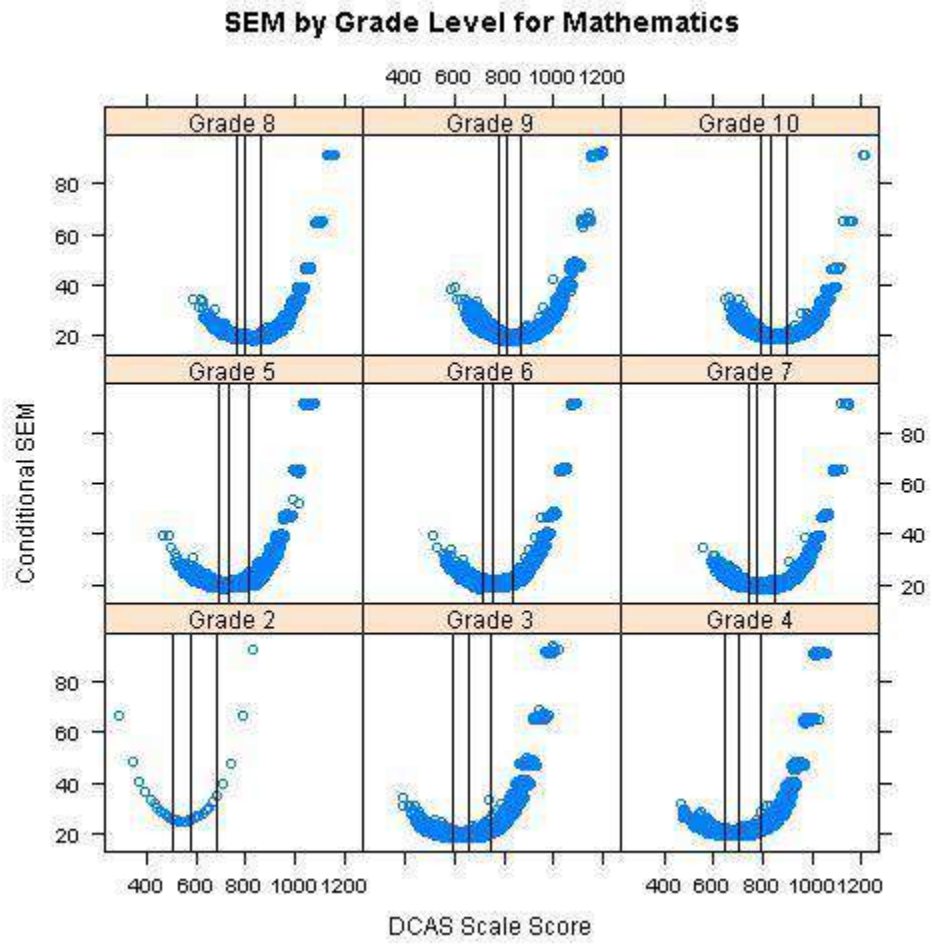


Figure 2:
Conditional Standard Errors of Measurement (Reading), DCAS Spring Window

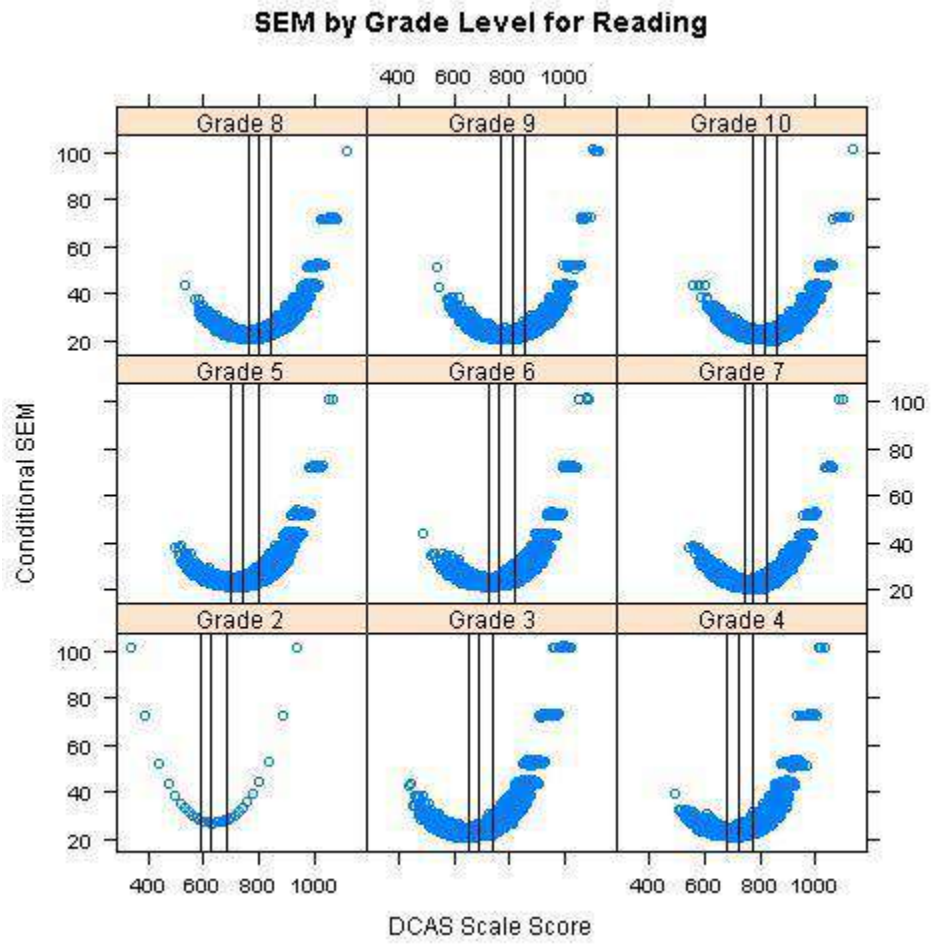


Figure 3:
Conditional Standard Errors of Measurement (Science), DCAS Spring Window

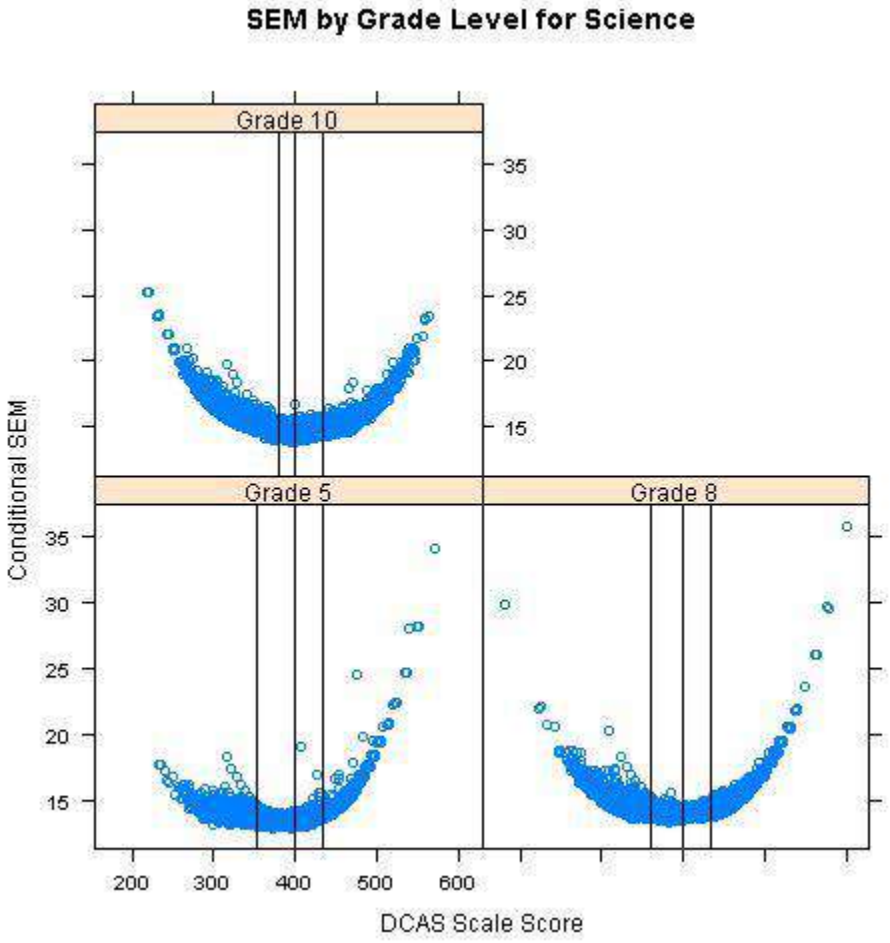
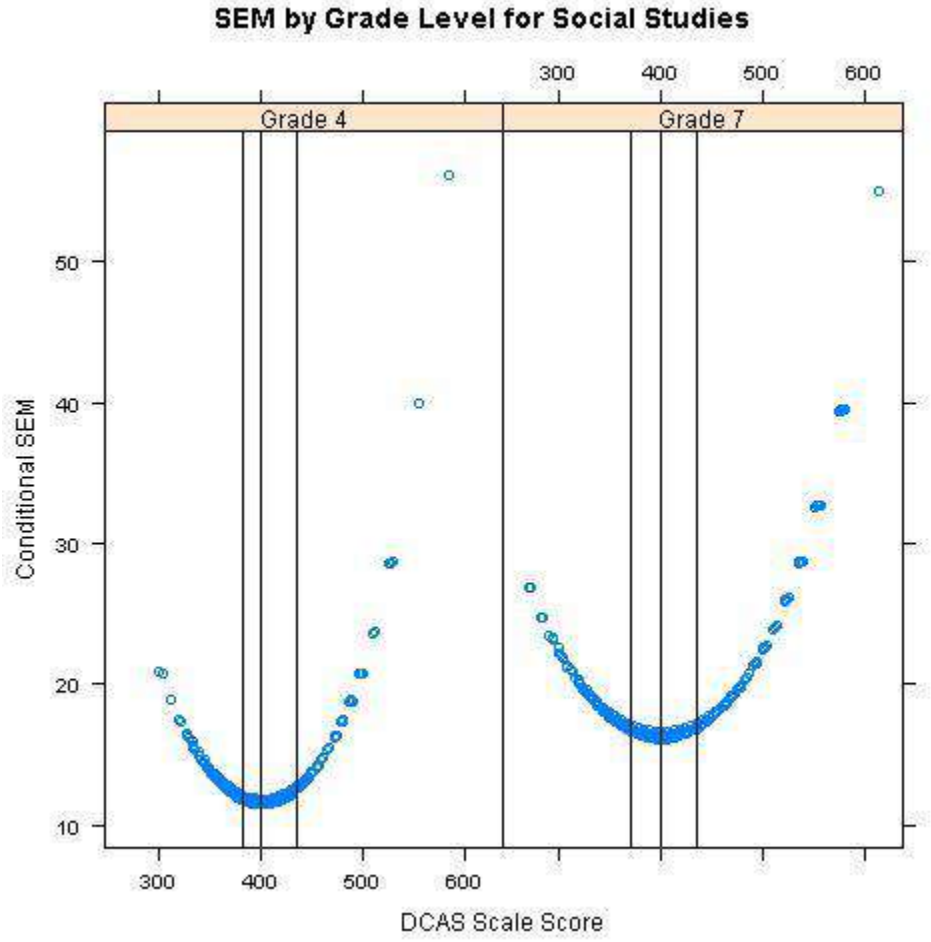


Figure 4:
Conditional Standard Errors of Measurement (Social Studies), DCAS Spring Window



The standard error curves suggest that students are measured with a very high degree of precision. In fact, the typical standard error for all tests approaches the theoretical lower limit for each test. However, we do observe larger standard errors at ends, especially the higher ends of the score distribution. This occurs because the current item pools lack items that are better targeted toward these high-achieving individuals. Content experts use this information to consider how to further target and populate item pools.

Tables 4–9 show the mean standard error of measurement of accountability scores for students scoring within each of the DCAS performance levels by grade and test window. Similarly, Tables 10–13 provide the standard deviation of scale score and mean standard error of measurement by grade and test window in all subjects.

Table 4: Reading Mean Standard Error of Measurement of Accountability Scores, DCAS Fall Window

Grade	Well Below	Below Standard	Meets Standard	Advanced
3	23.8	22.3	23.4	28.0
4	24.3	23.5	24.5	29.1
5	24.7	23.6	25.0	30.0
6	25.3	23.8	24.0	27.3
7	24.7	22.8	23.3	27.3
8	24.2	23.1	24.2	29.5
9	24.2	22.2	22.9	27.2
10	24.8	22.8	23.1	26.0

Table 5: Mathematics Mean Standard Error of Measurement of Accountability Scores, DCAS Fall Window

Grade	Well Below	Below Standard	Meets Standard	Advanced
3	20.5	19.7	19.8	23.4
4	21.2	19.9	20.1	23.5
5	20.2	19.1	19.0	20.9
6	22.3	19.9	19.3	22.0
7	20.5	19.1	18.9	20.4
8	20.6	19.2	18.9	20.7
9	20.9	19.4	19.1	20.5
10	20.9	19.1	18.6	20.0

Table 6: Reading Mean Standard Error of Measurement of Accountability Scores, DCAS Spring Window

Grade	Well Below	Below Standard	Meets Standard	Advanced
2	31.5	27.3	27.1	38.7
3	23.5	22.9	24.2	32.5
4	23.8	23.1	24.1	29.8
5	24.2	23.6	24.7	30.4
6	24.2	23.5	24.6	30.7
7	24.1	23.0	23.8	28.7
8	23.6	23.0	24.1	29.6
9	23.8	22.6	23.4	28.5
10	24.2	22.5	23.1	27.4

Table 7: Mathematics Mean Standard Error of Measurement of Accountability Scores, DCAS Spring Window

Grade	Well Below	Below Standard	Meets Standard	Advanced
2	28.5	25.1	29.7	58.7
3	19.9	19.1	19.7	27.3
4	20.7	19.9	20.4	28.5
5	20.6	19.6	20.0	27.6
6	21.3	20.1	20.6	28.4
7	20.3	19.2	19.1	22.3
8	20.6	19.2	18.9	23.3
9	21.6	19.8	19.3	22.5
10	21.9	20.1	19.2	21.9

Table 8: Science Mean Standard Error of Measurement of Scale Scores, DCAS Spring Window

Grade	Well Below	Below Standard	Meets Standard	Advanced
5	14.0	13.5	13.5	15.4
8	14.8	14.0	14.2	15.4
10	15.6	14.8	14.7	15.6

Table 9: Social Studies Mean Standard Error of Measurement of Scale Scores, DCAS Spring Window

Grade	Well Below	Below Standard	Meets Standard	Advanced
4	12.7	11.8	11.9	14.5
7	18.0	16.5	16.6	19.8

The test reliability information is also provided at the No Child Left Behind (NCLB) subgroup level. Appendix A of this volume provides the marginal reliability of the accountability scores disaggregated by subject and grade.

Table 10: Standard Deviation and Mean of SEM of Accountability Scores in Reading

Grade	Fall Window		Spring Window	
	Standard Deviation	Mean	Standard Deviation	Mean
2	-	-	92.8	33.1
3	78.1	24.1	77.8	27.7
4	72.1	25.0	64.9	26.3
5	73.2	25.5	67.1	27.0
6	65.2	24.9	65.6	26.6
7	69.3	24.5	63.6	25.7
8	76.8	25.4	67.7	26.3
9	76.8	24.2	71.6	25.5
10	68.2	24.3	63.6	25.4

Table 11: Standard Deviation and Mean SEM of Accountability Scores in Mathematics

Grade	Fall Window		Spring Window	
	Standard Deviation	SEM	Standard Deviation	SEM
2			87.7	36.6
3	69.2	20.2	79.4	21.7
4	65.9	20.6	78.4	22.4
5	64.3	19.7	74.4	22.1
6	62.7	20.8	76.5	22.7
7	59.1	19.6	64.8	20.0
8	61.8	19.8	68.3	20.6
9	65.8	20.0	71.3	20.8
10	60.2	19.6	60.7	20.3

Table 12: Standard Deviation and Mean SEM of Scale Scores in Science

Grade	Spring Window	
	Standard Deviation	SEM
5	42.3	13.9
8	46.6	14.5
10	48.6	15.2

Table 13: Standard Deviation and Mean SEM of Scale Scores in Social Studies

Grade	Spring Window	
	Standard Deviation	SEM
4	30.3	12.6
7	49.2	17.9

3.2 RELIABILITY OF ACHIEVEMENT CLASSIFICATION

Student performance on the DCAS is also reported in four performance categories for accountability score: Well Below Standard, Below Standard, Meets Standard, and Advanced. The standard-setting technical report provides detailed information about the standard-setting process, methodology, and results. These cut scores are used to classify student accountability scores into different performance levels.

The precision of classifying students as either proficient or not proficient (i.e., misclassification probabilities) is reported in Volume 1, Section 7.2.

3.3 REPORTING CATEGORY RELIABILITY AND PRECISION AT CUT SCORES

It is not sufficient to report measurement precision at the total score level only. We also need to provide evidence of precision for subscores. To evaluate the precision with which the content standards are measured using the subscale scores, we report a band of "indeterminacy" around the cut score for meeting standard within each of the subcontent areas, also referred to as reporting categories.

Although the performance standards were set based on the total accountability score, the same cut score can be projected onto each reporting category to evaluate a student's performance on that category since the subscores are on the same scale as the total accountability score. It is important to note that this cut score is used here only to provide some contextual evidence of subscore precision, not for reporting purposes.

The band of indeterminacy around the cut score of Meeting Standard for the reporting subcategory was created using the standard errors of the strand scores, not the standard errors derived from the total test score. These standard errors are, in general, larger than the standard errors for the overall test because they are derived on the basis of a small subset of items. Hence, we expect less precision in the subscores than in the test overall as a function of the smaller number of items.

Consequently, the projection using the Meets Standard cut onto each scale for strand scores divides the strand scores into three ranges: scores clearly below the cut (Below), scores that cannot be statistically distinguished from the cut (Near), and scores clearly above the cut (Above). The band of indeterminacy surrounding the cut score to create the Near classification is

$$near = \theta_j \pm se(\theta_j),$$

where θ_j is the proficient cut score of the *j*th reporting category and $se(\theta_j)$ is the standard error associated around this cut. Students are assigned to one of the three categories within each strand

using this band of indeterminacy. Students whose scores are within the band are placed into the Near category, students whose scores are above this range are classified as Above, and students whose scores are below this range are classified as Below.

For the DCAS, one scoring rule is that a reporting category must have at least eight items; otherwise, the strand scores are suppressed.

Tables 14–17 present the bandwidth for the Meets Standard category by subject, grade, and reporting category. For example, as presented in Table 14, grade 3 mathematics numeric reasoning has a bandwidth of 632–686 for the performance level of Meets Standard. This bandwidth was obtained by adding and subtracting the average standard error in the given reporting category (27) from the cuts for Meets Standard (659). A score of 635 is classified as Near, a score of 625 is classified as Below, and a score of 690 is classified as Above. Note that standard errors derived from reporting categories with a test length of fewer than eight items are unstable; therefore, their bandwidths are not reported in these tables.

Table 14: Bandwidth by Reporting Category, Mathematics, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Width (SE)	Min # of Items	Max # of Items
2	Numeric Reasoning	526–628 (51)	17	17
	Algebraic Reasoning	–	7	7
	Geometric Reasoning	–	4	4
	Quantitative Reasoning	–	2	2
3	Numeric Reasoning	632–686 (27)	30	33
	Algebraic Reasoning	605–713 (54)	8	9
	Geometric Reasoning	610–708 (49)	9	10
	Quantitative Reasoning	–	1	3
4	Numeric Reasoning	671–729 (29)	24	27
	Algebraic Reasoning	645–755 (55)	8	9
	Geometric Reasoning	655–745 (45)	12	14
	Quantitative Reasoning	–	4	5
5	Numeric Reasoning	702–762 (30)	22	24
	Algebraic Reasoning	680–784 (52)	8	10
	Geometric Reasoning	683–781 (49)	8	10
	Quantitative Reasoning	677–787 (55)	8	9
6	Numeric Reasoning	728–786 (29)	24	27
	Algebraic Reasoning	700–814 (57)	8	9
	Geometric Reasoning	704–810 (53)	8	9
	Quantitative Reasoning	703–811 (54)	8	9
7	Numeric Reasoning	746–812 (33)	16	18
	Algebraic Reasoning	742–816 (37)	15	17

Grade	Reporting Category	Width (SE)	Min # of Items	Max # of Items
	Geometric Reasoning	731–827 (48)	9	10
	Quantitative Reasoning	733–825 (46)	8	10
8	Numeric Reasoning	752–848 (48)	9	10
	Algebraic Reasoning	770–830 (30)	23	25
	Geometric Reasoning	751–849 (49)	8	9
	Quantitative Reasoning	752–848 (48)	8	9
9	Numeric Reasoning	–	2	3
	Algebraic Reasoning	788–836 (24)	35	38
	Geometric Reasoning	–	2	3
	Quantitative Reasoning	764–860 (48)	9	10
10	Numeric Reasoning	–	2	3
	Algebraic Reasoning	796–864 (34)	18	22
	Geometric Reasoning	799–861 (31)	20	21
	Quantitative Reasoning	779–881 (51)	8	9

Table 15: Bandwidth by Reporting Category, Reading, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Width (SE)	Min # of Items	Max # of Items
2	Comprehension	584–662(39)	22	22
	Literary Text	557–689(66)	8	8
3	Comprehension	659–721(31)	30	40
	Literary Text	641–739(49)	10	20
4	Comprehension	692–750(29)	30	40
	Literary Text	673–769(48)	10	20
5	Comprehension	709–769(30)	30	40
	Literary Text	690–788(49)	10	20
6	Comprehension	729–787(29)	30	40
	Literary Text	707–809(51)	10	20
7	Comprehension	747–805(29)	30	40
	Literary Text	729–823(47)	10	20
8	Comprehension	770–830(30)	30	40
	Literary Text	752–848(48)	10	20
9	Comprehension	782–840(29)	30	40
	Literary Text	763–859(48)	10	20

Grade	Reporting Category	Width (SE)	Min # of Items	Max # of Items
10	Comprehension	792–848(28)	30	40
	Literary Text	768–872(52)	10	20

Table 16: Bandwidth by Reporting Category, Science, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Width (SE)	Min # of Items	Max # of Items
5	Earth Science	372–428 (28)	10	14
	Life Science	377–423 (23)	17	21
	Physical Science	376–424 (24)	15	19
8	Earth Science	372–428 (28)	10	16
	Life Science	377–423 (23)	15	21
	Physical Science	374–426 (26)	12	18
10	Earth Science	363–437 (37)	8	12
	Life Science	375–425 (25)	17	21
	Physical Science	376–424 (24)	17	21

Table 17: Bandwidth by Reporting Category, Social Studies, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Width (SE)	Min # of Items	Max # of Items
4	Civics	373–427 (27)	11	13
	Economics	373–427 (27)	11	12
	Geography	374–426 (26)	11	12
	History	373–427 (27)	11	12
7	Civics	358–442 (42)	11	14
	Economics	362–438 (38)	11	12
	Geography	363–437 (37)	11	12
	History	362–438 (38)	11	12

Tables 18–21 present the distribution of students' subscores by each reporting category. Subscales with a minimum test length of fewer than eight items provide unstable estimates and are not reported in these tables. Some of the key observations from these tables include the following:

- For mathematics (Table 18), between 9% and 22% of the students are at the Below level. Grade 2 has the smallest percentage of students who are at the Below level.

- For reading (Table 19), between 14% and 23% of the students are at the Below level in Reading Comprehension, and between 12% and 20% of students are at the Below level in Literary Text. Throughout all grades, the percentage of students who are at the Above level is higher in Comprehension than in Literary Text.
- For science (Table 20), the percentage of students at the Below level ranges from 28% (Earth Science in both grades 5 and 8) to 35% (Life Science, grade 10). In grade 10, more students are at the Below level than the Above level.
- For social studies (Table 21), almost half of the grade 4 students are at the Near level in Economics, Geography, and History. The percentages are similar in Economics and Geography for grade 7 students. The percentages at the Above level are higher as compared to the percentages at the Below level in both grades.

Table 18: Number and Percent Distribution of Students by Reporting Category, Mathematics, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Below	%	Near	%	Above	%
2	Numeric Reasoning	896	9	3247	31	6258	60
	Algebraic Reasoning	407		4055		5939	
	Geometric Reasoning	490		5348		4563	
	Quantitative Reasoning	690		9711			
3	Numeric Reasoning	2025	20	2263	22	5874	58
	Algebraic Reasoning	1324	13	4345	43	4493	44
	Geometric Reasoning	1519	15	3991	39	4652	46
	Quantitative Reasoning	763		9282		117	
4	Numeric Reasoning	1589	16	2436	24	6027	60
	Algebraic Reasoning	1139	11	4410	44	4503	45
	Geometric Reasoning	1431	14	4236	42	4385	44
	Quantitative Reasoning	1220		5113		3719	
5	Numeric Reasoning	1843	18	3121	30	5295	52
	Algebraic Reasoning	1344	13	4574	45	4341	42
	Geometric Reasoning	1587	15	4778	47	3894	38
	Quantitative Reasoning	1158	11	4794	47	4307	42
6	Numeric Reasoning	2059	21	2962	30	4896	49
	Algebraic Reasoning	1269	13	4376	44	4272	43
	Geometric Reasoning	1566	16	5174	52	3177	32
	Quantitative Reasoning	1448	15	4946	50	3523	36
7	Numeric Reasoning	1760	18	3649	37	4548	46
	Algebraic Reasoning	1875	19	3824	38	4258	43
	Geometric Reasoning	1572	16	5161	52	3224	32
	Quantitative Reasoning	1885	19	4342	44	3730	37

Grade	Reporting Category	Below	%	Near	%	Above	%
8	Numeric Reasoning	1252	13	4705	47	4043	40
	Algebraic Reasoning	1889	19	3134	31	4977	50
	Geometric Reasoning	1353	14	4957	50	3690	37
	Quantitative Reasoning	1358	14	4963	50	3679	37
9	Numeric Reasoning	1217		6885		2993	
	Algebraic Reasoning	2440	22	2896	26	5759	52
	Geometric Reasoning	1494		6170		3431	
	Quantitative Reasoning	1697	15	5397	49	4001	36
10	Numeric Reasoning	1537		5907		1709	
	Algebraic Reasoning	1525	17	3461	38	4167	46
	Geometric Reasoning	1448	16	3665	40	4040	44
	Quantitative Reasoning	1428	16	4480	49	3245	35

Table 19: Number and Percent Distribution of Students by Reporting Category, Reading, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Below	%	Near	%	Above	%
2	Comprehension	2076	20	2779	27	5511	53
	Literary Text	2068	20	3968	38	4330	42
3	Comprehension	1964	19	2816	28	5359	53
	Literary Text	1390	14	4225	42	4524	45
4	Comprehension	1555	16	2982	30	5494	55
	Literary Text	1267	13	4256	42	4508	45
5	Comprehension	1408	14	2779	27	6033	59
	Literary Text	1186	12	4034	39	5000	49
6	Comprehension	1580	16	2894	29	5414	55
	Literary Text	1207	12	4505	46	4176	42
7	Comprehension	1773	18	2801	28	5355	54
	Literary Text	1311	13	4609	46	4009	40
8	Comprehension	1741	17	3006	30	5208	52
	Literary Text	1361	14	4328	43	4266	43
9	Comprehension	2572	23	3212	29	5204	47
	Literary Text	1741	16	4752	43	4495	41
10	Comprehension	1414	16	2397	26	5298	58
	Literary Text	1212	13	4402	48	3495	38

Table 20: Number and Percent Distribution of Students by Reporting Category, Science, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Below	%	Near	%	Above	%
5	Earth Science	2867	28	4643	46	2642	26
	Life Science	2991	29	3930	39	3231	32
	Physical Science	3213	32	4041	40	2898	29
8	Earth Science	2690	28	4003	42	2921	30
	Life Science	2950	31	3451	36	3213	33
	Physical Science	2832	29	3921	41	2861	30
10	Earth Science	2538	29	4525	51	1736	20
	Life Science	3085	35	3244	37	2470	28
	Physical Science	2958	34	3602	41	2239	25

Table 21: Number and Percent Distribution of Students by Reporting Category, Social Studies, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Below	%	Near	%	Above	%
4	Civics	1351	14	4449	45	4132	42
	Economics	1335	13	5194	52	3403	34
	Geography	1917	19	5332	54	2683	27
	History	1622	16	4914	49	3396	34
7	Civics	1300	13	3783	39	4584	47
	Economics	1638	17	4629	48	3400	35
	Geography	2229	23	4662	48	2776	29
	History	2363	24	4321	45	2983	31

4. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills covered by DCAS are representative of the prioritized content standards of the larger knowledge domain. We describe the content standards for DCAS and the test development process that maps DCAS tests to these prioritized standards.

4.1 CONTENT STANDARDS

In 1995, Delaware adopted content standards for English language arts (ELA), mathematics, science, and social studies. Immediately, the Delaware Department of Education (DDOE) began developing the Delaware Student Testing Program (DSTP), and in 1998, the state legislature passed laws (Delaware Code, Title 14, §151 and §152) that made the DSTP the official measure of student progress toward the Delaware Content Standards and a major measurement tool for Delaware's accountability system. In the 2010–2011 school year, DDOE implemented a new online assessment system, the Delaware Comprehensive Assessment System (DCAS), which replaced the DSTP paper-and-pencil test. The DCAS assessments that were administered in the 2013–2014 school year follow the same test specification and framework as those administered in previous school years (2010–2011, 2011–2012, and 2012–2013).

The terminology used in referring to the content standard varies by subjects. For example, the highest level in the content hierarchy is referred to as "standards" in mathematics, reading, and social studies, while "reporting category" is used in science. For DCAS, these levels are also used for reporting purposes. In addition, we use the terms "reporting category," "standards," and "subscale" interchangeably throughout this document. The content standards/reporting category for each subject for the 2013–2014 DCAS are presented in Table 22.

Table 22: Standards/Reporting Category for DCAS 2013–2014

Subject	Content Standards/Reporting Category
Reading Grades 2–10	Reading Comprehension
	Literary Text
Mathematics Grades 2–10	Numeric Reasoning
	Algebraic Reasoning
	Geometric Reasoning
	Quantitative Reasoning
Science Grades 5, 8, 10	Life Science
	Earth Science
	Physical Science
Social Studies Grades 4, 7	Civics
	Economics
	Geography
	History

4.2 TEST SPECIFICATIONS

Test specifications are blueprints developed to ensure that the test and the items are aligned to the prioritized standards that they are intended to measure. Multiple-choice (MC) and machine-scored constructed-response (MSCR) items are used in the DCAS. For CAT, test specifications (blueprints) stipulate the minimum and maximum number of operational on-grade items (for accountability scores) and both on-grade and off-grade items (for instructional scores) that must be administered for a given test. For the linear tests, the blueprint specifies the percentage of operational items that must be administered. The blueprints also include the minimum and maximum number of on- and off-grade items for each of the reporting categories and constraints on selecting items for the depth of knowledge (DOK) levels in reading. The minimum and maximum number of items by grade and subject and other details on the blueprint are presented in Volume 2, Section 1.1, of the 2011 DCAS Technical Report.

4.3 TEST DEVELOPMENT

The items used in the embedded field test (EFT) of spring 2014 came from various sources. Note that only mathematics and social studies tests included field-test items. In mathematics, the items came from the following sources, with grade(s) in parentheses:

- Development from prior year that had no field-test slots available during spring 2013 field-test window (grades 3–6)
- Items that did not survive spring 2013 data review and were re-field-tested with or without edits (grades 3–10)
- Items originally field-tested for end-of-course (EOC) tests but never used on operational forms and those re-field-tested at Common Core grades (grades 6–10)
- EOC items that could not be field-tested due to limited field-test slots (grade 10)

The EFT items used in spring 2014 test window came from two sources: newly developed items and some items that were re-field-tested for a variety of reasons. To develop new items, content experts from the contractor, AIR, DDOE, and Delaware educators worked together to review and select items for the field test and to then add new items to the item pool. This section provides an overview of the test development process. See Volume II, Section 2, of the 2010–2011 DCAS Technical Report for a more detailed description of the new item development procedure, including the criteria used in passage selection and item writing and the quality control and review process.

Development of New Items

New items are originally developed by content specialists at AIR. All newly developed items are reviewed and revised as needed and pass through an extensive review process, including various AIR content and editorial reviews, a DDOE content review, a bias review by the Fairness and Sensitivity Committee, and a content review by the Content Advisory Committee (CAC) before they can be included in the field-test pools.

Items that survive AIR internal reviews are sent to the DDOE for review. DDOE content and assessment experts review and rate each item and render a decision as to its fate: items are accepted, sent back for revisions and then reviewed again, or rejected. AIR content experts and DDOE staff discuss suggested revisions and come to agreement, after which changes are made.

Following the completion of the AIR and DDOE internal reviews, the items are reviewed by two Delaware committees: the Fairness and Sensitivity Committee and the CAC.

The CAC consists of Delaware grade-appropriate classroom teachers for each subject area and occasional content experts in higher education or industry. The primary responsibility of the committee members is to ensure that the items are based on defensible content and are free from such flaws as inappropriate readability level, ambiguity, multiple answer keys, and unclear instructions. These items are approved, approved with modifications, revised by AIR under DDOE direction, or rejected.

Items that have passed through CAC review are then reviewed by the Fairness and Sensitivity Committee. This committee specifically reviews items for potential bias and controversial content and attempts to identify any items that are likely to present bias for specific groups of Delaware students. The Fairness and Sensitivity Committee comprises Delaware educators who are selected to ensure geographic and ethnic diversity. The committee ensures that items

- present racial, ethnic, and cultural groups in a positive light;
- do not contain controversial, offensive, or potentially upsetting content;
- avoid content familiar only to specific groups of students because of race or ethnicity, class, or geographic location;
- aid in the elimination of stereotypes; and
- avoid words or phrases that have multiple meanings.

DDOE and AIR reject or edit items based on Fairness and Sensitivity Committee recommendations. Items that are approved by both of these committees will advance to be field-tested.

After the field test is completed, members of the Rubric Validation Committee review the responses provided for every MSCR item and either approve the scoring rubric or suggest a revised score based on their interpretation of the item task and the rubric. More details on the review process of these various committees are provided in Volume 2, Section 2, of the 2011 DCAS Technical Report.

4.4 ALIGNMENT OF DCAS ITEM BANKS TO THE CONTENT STANDARDS AND BENCHMARKS

Item alignment is an integral component of test development. An alignment study reviews and determines the degree to which the test and the standards set are in agreement with and support student learning of intended expectations. To maintain objectivity, the alignment study that evaluates the alignment of DCAS item banks to the content standards was completed by an independent contractor, Dr. Norman Webb, a nationally recognized expert in alignment of state

assessment programs. Note that the alignment study was conducted during the summer of 2010, which was prior to the first operational test window in the 2010–2011 school year.

4.4.1 Summary of Webb Alignment Study

A detailed description of the alignment study is presented in Volume 2, Section 2.4, of the 2011 DCAS Technical Report. Additionally, the independent contractor that performed this analysis provided the DDOE with a separate report on the findings. In general, the item banks in science and social studies are generally aligned to the corresponding standards per grade. In mathematics, the alignment for grades 3–8 is good. In reading, the alignment needs some improvement.

5. EVIDENCE ON INTERNAL STRUCTURE

In this section, we explore the internal structure of the assessment using the scores provided at the strand level. The relationship of the subscores is but one indicator of the test dimensionality.

In mathematics, there are four reporting categories per grade: Numeric Reasoning, Geometric Reasoning, Algebraic Reasoning, and Quantitative Reasoning. In reading, there are two reporting categories: Reading Comprehension and Literary Text. In science, there are three reporting categories: Life Science, Earth Science, and Physical Science. In social studies, there are four reporting categories: Economics, Civics, Geography, and History.

The DCAS strand scores are based on a subset of items under each category and reported on the same scale as the total test score. Evidence is needed to verify that these strand scores provide different and useful information for student achievement.

It may not be reasonable to expect that the strand scores are completely orthogonal—this would suggest that there are no relationships among strand scores and would make justification of a unidimensional IRT model difficult, although we could then easily justify reporting these separate scores. On the other hand, if they are perfectly correlated, we could justify a unidimensional model, but we could not justify the reporting of separate scores.

In some cases, it may be useful to propose a second-order factor model and compare this to a purely unidimensional model. For instance, we might assume that a general mathematics construct (first factor) with four strands (second factor) and that the items load onto the strands they intend to measure; we would then compare this to a unidimensional model where all items load onto a single factor.

However, the fit statistics typically extracted from confirmatory factor models often reveal significant differences between these two approaches that are a function of total sample size. For this reason, we provide a more pragmatic approach.

The data below are the observed correlations between the strand scores. Because each standard is measured with a small number of items, the standard errors of the observed scores within each standard are typically larger than the standard error of the total test score. A second complicating factor is that these correlations are expected to be lower than the correlations based on true scores. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. However, the correction for measurement, in some cases, may produce a correlation exceeding 1.0. Consequently, only the observed score correlations are presented.

5.1 CORRELATIONS AMONG STRAND SCORES

Tables 23–26 present the correlation matrix of the strand-level scores for each subject area. Because of unstable estimates, the minimum number of eight items under each reporting category is used for analysis in this report. The minimum and maximum numbers of items by standard are based on test specifications.

In mathematics, the correlations among the four subscales (standards) range from 0.54 to 0.75. For reading, the correlations between Reading Comprehension and Literary Text range from 0.65 to 0.69. For science, the correlations between the three subscales—Earth Science, Life Science,

and Physical Science—vary between 0.65 and 0.76 across grades. For social studies, the correlations between the subscales of Economics, Civics, History, and Geography fall between 0.47 and 0.64.

In some cases, these correlations seem low. However, as previously noted, the correlations are subject to the large amount of measurement error at the strand level, given the limited number of items from which the scores are derived. Consequently, over-interpretation of these correlations (as either high or low) should be made cautiously.

Table 23: Correlation Matrix Among Reporting Categories, Mathematics, DCAS 2013–2014 Spring Window

Grade	Reporting Category	NR	AR	GR	QR	Min Number of Items	Max Number of Items
2	Numeric Reasoning (NR)	1				17	17
	Algebraic Reasoning (AR)	-	-			7	7
	Geometric Reasoning (GR)	-	-	-		4	4
	Quantitative Reasoning (QR)	-	-	-	-	2	2
3	Numeric Reasoning (NR)	1				30	33
	Algebraic Reasoning (AR)	0.75	1			8	9
	Geometric Reasoning (GR)	0.74	0.62	1		9	10
	Quantitative Reasoning (QR)	-	-	-	-	1	3
4	Numeric Reasoning (NR)	1				24	27
	Algebraic Reasoning (AR)	0.71	1			8	9
	Geometric Reasoning (GR)	0.71	0.59	1		12	14
	Quantitative Reasoning (QR)	-	-	-	-	4	5
5	Numeric Reasoning (NR)	1				22	24
	Algebraic Reasoning (AR)	0.73	1			8	10
	Geometric Reasoning (GR)	0.71	0.62	1		8	10
	Quantitative Reasoning (QR)	0.65	0.59	0.56	1	8	9
6	Numeric Reasoning (NR)	1				24	27
	Algebraic Reasoning (AR)	0.74	1			8	9
	Geometric Reasoning (GR)	0.63	0.56	1		8	9
	Quantitative Reasoning (QR)	0.71	0.65	0.54	1	8	9
7	Numeric Reasoning (NR)	1				16	18
	Algebraic Reasoning (AR)	0.74	1			15	17
	Geometric Reasoning (GR)	0.63	0.63	1		9	10
	Quantitative Reasoning (QR)	0.67	0.65	0.56	1	8	10
8	Numeric Reasoning (NR)	1				9	10

Grade	Reporting Category	NR	AR	GR	QR	Min Number of Items	Max Number of Items
	Algebraic Reasoning (AR)	0.74	1			23	25
	Geometric Reasoning (GR)	0.63	0.68	1		8	9
	Quantitative Reasoning (QR)	0.59	0.65	0.57	1	8	9
9	Numeric Reasoning (NR)	-				2	3
	Algebraic Reasoning (AR)	-	1			35	38
	Geometric Reasoning (GR)	-	-	-		2	3
	Quantitative Reasoning (QR)	-	0.70	-	1	9	10
10	Numeric Reasoning (NR)	-				2	3
	Algebraic Reasoning (AR)	-	1			18	22
	Geometric Reasoning (GR)	-	0.72	1		20	21
	Quantitative Reasoning (QR)	-	0.62	0.62	1	8	9

Table 24: Correlation Matrix Among Reporting Categories, Reading, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Reading Comprehension	Literary Text	Min Number of Items	Max Number of Items
2	Reading Comprehension	1		22	22
	Literary Text	0.74	-	8	8
3	Reading Comprehension	1		30	40
	Literary Text	0.71	1	10	20
4	Reading Comprehension	1		30	40
	Literary Text	0.66	1	10	20
5	Reading Comprehension	1		30	40
	Literary Text	0.69	1	10	20
6	Reading Comprehension	1		30	40
	Literary Text	0.67	1	10	20
7	Reading Comprehension	1		30	40
	Literary Text	0.66	1	10	20
8	Reading Comprehension	1		30	40
	Literary Text	0.68	1	10	20
9	Reading Comprehension	1		30	40
	Literary Text	0.68	1	10	20
10	Reading Comprehension	1		30	40
	Literary Text	0.65	1	10	20

Table 25: Correlation Matrix Among Reporting Categories, Science, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Earth Science	Life Science	Physical Science	Min Number of Items	Max Number of Items
5	Earth Science	1			10	14
	Life Science	0.68	1		17	21
	Physical Science	0.66	0.72	1	15	19
8	Earth Science	1			10	16
	Life Science	0.71	1		15	21
	Physical Science	0.70	0.75	1	12	18
10	Earth Science	1			8	12
	Life Science	0.65	1		17	21
	Physical Science	0.65	0.76	1	17	21

Table 26: Correlation Matrix Among Reporting Categories, Social Studies, DCAS 2013–2014 Spring Window

Grade	Reporting Category	Civics	Economics	Geography	History	Min Number of Items	Max Number of Items
4	Civics	1				11	13
	Economics	0.53	1			11	12
	Geography	0.47	0.47	1		11	12
	History	0.55	0.55	0.50	1	11	12
7	Civics	1				11	14
	Economics	0.60	1			11	12
	Geography	0.55	0.57	1		11	12
	History	0.60	0.64	0.60	1	11	12

6. EVIDENCE OF COMPARABILITY

When multiple test forms of paper-and-pencil versions are constructed, it is important to provide evidence of comparability between the forms as well as with the adaptive testing. With the CAT, we must be concerned with comparability of scores as well as comparability of content. If the content between forms varies, then it will be difficult to justify score comparability.

Student scores should not depend on the mode of administration or the type of test form. DCAS is an online assessment. To improve the accessibility of the statewide assessment, alternate assessments are provided for students with special needs. Thus, the comparability of scores obtained via alternate means of administration must be established and evaluated.

6.1 MATCH-WITH-TEST BLUEPRINTS FOR BOTH PAPER-AND-PENCIL AND ONLINE TESTS

For the 2013–2014 DCAS, the paper-and-pencil version of the tests was the same as those administered previous years. Those tests were developed according to the same test specifications used for the online adaptive tests. The match-to-blueprint is therefore 100%, given that the paper forms are developed to match the specification. In this section, evidence for both is provided. The procedures used to establish comparable fixed forms are provided in Volume 2, Test Development, of the 2011 DCAS Technical Report.

Matching the blueprint is critically important for online adaptive tests because students took different sets of items—the adaptive algorithm chooses the items to be presented to the student while the student is taking the test. Therefore, it is important to determine whether each adaptive test meets the specified blueprints.

If the priority of the algorithm were to match ability, then it would construct test forms that varied in terms of content characteristics. However, the algorithm used for the DCAS is explicitly designed to operate and prioritize item selection according to two criteria:

1. Match test specifications
2. Match student ability

A complete description of the algorithm can be found in Volume 1, Section 3.3, of the 2013–2014 DCAS Annual Technical Report.

Table 27 presents the percentage of CATs meeting the blueprint for all test opportunities and subjects using the operational data based on the fixed length of 50 items. These results are consistent with the results obtained from simulation studies, which are summarized above in Section 3.2 of the 2013–2014 DCAS Annual Technical Report. The blueprint match is 100% in all grades and subjects. Note that the test blueprints are available in Volume 2, Appendix B, of the 2011 DCAS Technical Report. These percentages are computed at the reporting category level of the test specifications.

Table 27: Percentage of Online DCAS Tests Meeting Blueprint, 2013–2014

Grade	Mathematics				Reading				Science
	All OPPs* combined	OPP* 1	OPP 2	OPP 3	All OPPs combined	OPP 1	OPP 2	OPP 3	OPP 1
2	100	100			100	100			
3	100	100	100	100	100	100	100	100	
4	100	100	100	100	100	100	100	100	
5	100	100	100	100	100	100	100	100	100
6	100	100	100	100	100	100	100	100	
7	100	100	100	100	100	100	100	100	
8	100	100	100	100	100	100	100	100	100
9	100	100	100	100	100	100	100	100	
10	100	100	100	100	100	100	100	100	100

*OPPs = Opportunities

6.2 Comparability of DCAS Test Scores Over Time

At the beginning of the DCAS implementation, all test items were initially calibrated and scaled using the Rasch model and Masters' partial credit model (PCM) (Masters, 1982) through an independent field test in spring 2010. Any items added to the pool later via embedded field-testing are calibrated by anchoring on operational item parameters from the bank. Because test scores must be immediately provided to students upon conclusion of a test, pre-equating is used for the DCAS assessment, and scores are provided given the existing item parameters.

However, with the CAT, each individual form is theoretically unique, though there are some overlapping items as tests are tailored to student ability, and the mean difficulty of the form varies according to how students respond to the test items. However, all selected item parameters used to construct scores are drawn from a single item bank and are on the same scale of measurement. This ensures that scores between students are also on the same scale. The score comparability exists based on that all items are calibrated on the same scale, and the item selection process is based on the same blueprint.

6.3 Comparability of Computer-Adaptive and Paper-and-Pencil Test Scores

Linear test forms for paper-and-pencil administration are offered as a special accommodation for students with special needs. These fixed forms align to the same test specifications as the CAT, and use the same item parameters for scoring. However, without an online system, MSCR items cannot be administered with paper-and-pencil testing. This is the only difference between the two versions. Standard error plots overlaying the adaptive tests with the paper-and-pencil tests for reading, mathematics, and science are presented in Appendix B. These plots show that, as expected, each test has larger standard errors on the upper and lower ends of the scale scores.

6.4 TRANSLATION ACCURACY FROM ENGLISH TO SPANISH

DCAS items were originally developed in English. To meet the needs of some Hispanic students, test items are translated into Spanish by a professional language translation firm for the alternate version in mathematics, science, and social studies. Three steps were stipulated in the process of translating the items:

1. Translation: All content, including text-related graphics, are translated into Spanish.
2. Vendor quality control (QC): The translated item is reviewed and edited for content and/or style, if necessary, by a second translator who did not do the original translation. The English text in the graphics is translated at this phase.
3. Vendor sign-off: The translated item, including the translated text in graphics, the format, and the display, goes through one more review with final sign-off by the vendor.

At the QC and sign-off phases, the reviewers have a checklist to follow, which includes the following items:

- All text that needs to be translated has been translated.
- The graphics display properly and are well adapted to the translation.
- The fonts and upper/lower cases are consistent.
- The tables/charts are still aligned properly.
- The spacing and disposition of the text are the same as in the English version.

After vendor sign-off, AIR implements an internal review process for the translated items with various review levels. For example, in the Spanish Web Preview/Approval stage (the final review and approval level for the Spanish translation), if a discrepancy is identified, the item is moved back to the Spanish translation step. In other stages, the initial Spanish descriptions (tags) are entered for Spanish text-to-speech (TTS) by AIR bilingual professionals. Next, all Spanish content is reviewed for accuracy to ensure that each item accurately conveys the intent of the English text. If a discrepancy is identified, the item is held at this level for consultation with the Spanish vendor. The Spanish tags are reviewed and modified accordingly before translated items appear in the test.

7. FAIRNESS AND ACCESSIBILITY

7.1 FAIRNESS IN CONTENT

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. Universal design removes barriers to provide access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Test development specialists receive extensive training on the principles of universal design and apply these principles in the development of all test materials, including tasks, items, and manipulatives. In the review process, adherence to the principles of universal design is verified.

7.2 STATISTICAL FAIRNESS IN ITEM STATISTICS

The DCAS independent field test was conducted in spring 2010. Similarly, additional items were field-tested by embedding with the operational tests during the spring window in each of the past four years (spring 2011, spring 2012, spring 2013, and spring 2014). As discussed in Section 4.3 of this volume, all newly developed items pass through an extensive review process, including various AIR content and editorial reviews, a DDOE content review, a bias review by the Fairness and Sensitivity Committee, and a content review by the Content Advisory Committee (CAC) before they can be included in the field-test pools. Volume 2, Section 2.2 of the 2010–2011 DCAS Technical Report provides a detailed development and review process.

Following the field test in each of those administrations, differential item functioning (DIF) analyses were conducted for all field-tested items to detect potential item bias from a statistical perspective across major ethnic and gender groups. DIF analyses were performed for the following groups:

- Male/Female
- White/African-American
- Special education/No special education

Items are classified into three categories (A, B, or C) for DIF, ranging from no evidence of DIF to severe DIF according to the DIF classification convention illustrated in Table 28. Items are also categorized as positive DIF (i.e., +A, +B, or +C), signifying that the item favors the focal group (e.g., African-American/black, Hispanic, or female), or negative DIF (i.e., –A,–B, or–C), signifying that the item favors the reference group (e.g., white or male). Items are flagged if their DIF statistics fall into the "C" category for any group. A DIF classification of "C" indicates that the item shows significant differential item functioning and should be reviewed for potential content bias, differential validity, or other issues that may reduce item fairness, whether the DIF statistic favors the focal or the reference group.

Table 28: DIF Classification Rules

Dichotomous Items	
Category	**Rule**
C	$MH\chi^2$ is significant and $\lvert \hat{\Delta}_{MH} \rvert \geq 1.5$.
B	$MH\chi^2$ is significant and $\lvert \hat{\Delta}_{MH} \rvert < 1.5$.
A	$MH\chi^2$ is not significant.
Polytomous Items	
Category	**Rule**
C	$MH\chi^2$ is significant and $\lvert SMD \rvert / \lvert SD \rvert \geq .25$.
B	$MH\chi^2$ is significant and $\lvert SMD \rvert / \lvert SD \rvert < .25$.
A	$MH\chi^2$ is not significant.

A detailed description of the DIF analysis performed is presented in Volume 1, Section 5.2, of the 2013–2014 DCAS Annual Technical Report. The DIF statistics for each field-test item are presented in Volume 1, Appendix H, of the 2013–2014 DCAS Annual Technical Report.

Flagged items passed through a two-stage data review process—AIR internal review and DDOE review—before they were included in the final item pool for operation. The results from field-test analysis, including the number of items flagged and the results from item data review meeting, are also presented in Volume 1, Section 5, of the 2013–2014 DCAS Annual Technical Report.

Summary

Data presented in this report provide empirical evidence on internal structure validity, content validity, and reliability of test scores.

- Internal structural validity: Correlations among subscale scores within each content area in DCAS also achieve the expectations.
- Content validity: Data show that the match-to-blueprint rates are high, ensuring that content coverage on each form is consistent with test specifications.

- Reliability: The standard error curves show that students are measured with a very high degree of precision, although larger standard errors are observed at the higher ends of the score than at the lower end. Classification accuracy analysis also shows that students are classified as either proficient or not proficient with a high degree of certainty.

8. REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum.

Thompson, S., Johnstone, C., & Thurlow, M. (2002). *Universal design applied to large scale assessments.* NCEO Synthesis Report 44. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.