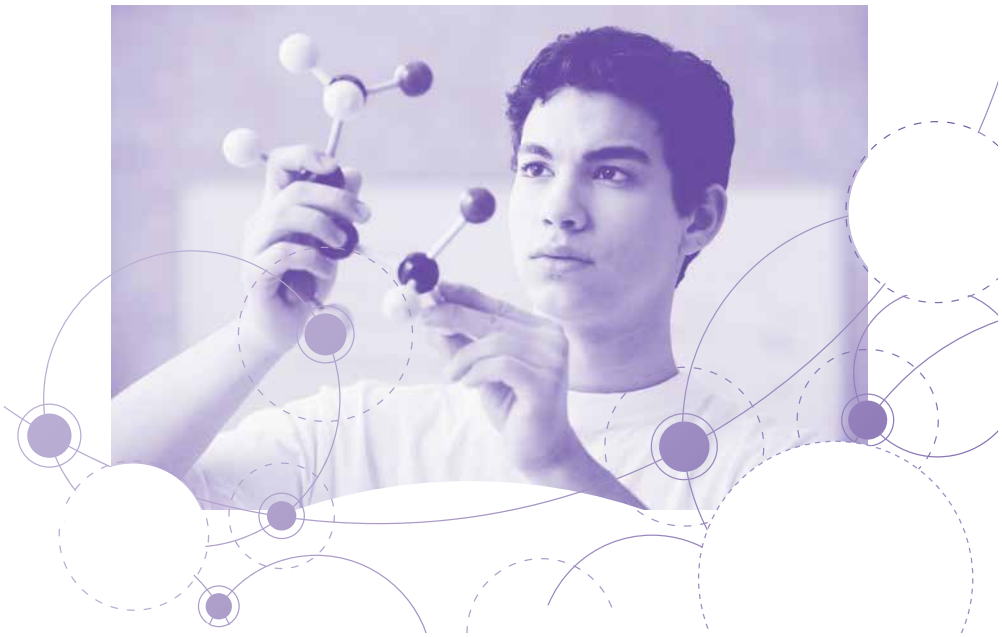


Summative Assessment

TECHNICAL BULLETIN #2: NORMS, SCORING, SCALING, AND PSYCHOMETRICS

ACT® Aspire®

discoveractaspire.org

ACT endorses the *Code of Fair Testing Practices in Education* and the *Code of Professional Responsibilities in Educational Measurement*, guides to the conduct of those involved in educational testing. ACT is committed to ensuring that each of its testing programs upholds the guidelines in each *Code*.

© 2014 by ACT, Inc. All rights reserved. ACT Aspire®, ACT®, ACT Explore®, ACT Plan®, EPAS®, and ACT NCRC® are registered trademarks, and ACT National Career Readiness Certificate™ is a trademark of ACT, Inc. 2631

Contents

Overview 1

Introduction 2

1 ACT Aspire Score Scale 3

ACT Aspire Scaling Philosophy 3

Scaling Study 3

Introduction to the Scaling Study Design 3

Construction of Scaling Tests 4

Data Collection Design 5

Creating the Vertical Scale 6

Evaluation of the Scales 16

Results 17

Scaling Test Raw Scores 17

Evaluating the On-the-Same-Scale Property 17

Evaluating the Constant CSEM Property 26

Growth Patterns 30

Scaling ACT Aspire Writing 35

2 ACT Aspire Scores 36

ACT Aspire Subject Scores 36

ACT Aspire Composite Score 36

ACT Aspire ELA Score 38

ACT Aspire STEM Score 38

Reporting Categories 39

Progress with Text Complexity 39

3 ACT Aspire Norms 40

4 EPAS® to ACT Aspire Concordance 61

Method 62

Evaluation of Results 62

Results 63

Evaluating Results 66

Summary 75

5 ACT Readiness Benchmarks 76

Introduction 76

ACT Readiness Benchmarks for English, Mathematics, Reading, and Science 76
Grades 8–10 77
Grades 3–7 78
ACT Readiness Benchmark for Writing 79
ACT Readiness Benchmarks for ELA and STEM 79
ACT Readiness Ranges 82
Readiness Ranges for Reporting Categories 82

6 ACT Aspire Growth 83
Predicted Score Paths 83
Student Progress Reports 84
Aggregate Progress Reports 84
Samples Used to Develop the Predicted Paths Used for 2014 Reporting 85
Statistical Model Used to Develop the Predicted Paths 86
Coverage Rates of the Predicted Paths 86
Limitations of the Predicted Paths 86
Student Growth Percentiles 87
ACT Aspire Gain Scores 87
Interpreting ACT Aspire Gain Scores 88
Growth-to-Standards Models with ACT Aspire 89
Measurement Error of Growth Scores 90
Aggregate Growth Scores for Research and Evaluation 90

7 Progress toward Career Readiness 91
Introduction 91
Link ACT NCRC Levels to the EPAS Composite Scores 93
Link the EPAS Composite Scores to ACT Aspire Composite Scores 94
Identify the ACT Aspire Composite Scores Corresponding to Each ACT NCRC Level . . . 95
Report Progress toward Career Readiness using ACT Aspire Cut Scores 96

8 ACT Aspire Reliability 97
Raw Score Reliability 98
Scale Score Reliability and Conditional Standard Error of Measurement 101

9 ACT Aspire Validity 103
Study 1: Comparison of ACT Explore and ACT Plan Scores to ACT Aspire Scores 104
Study 2: Comparison of State Assessment Scores to ACT Aspire Scores 119
Summary 134

10 ACT Aspire Mode Comparability Study 136
Introduction 136

Method and Results 137
Mode Effects for Raw Number-of-Points Scores on Common Items. 138
Comparisons of Scale Scores across Mode 155
Summary. 173

11 ACT Aspire Equating 174

Appendix A: The ACT with Constructed-Response Items. 176

Appendix B: EPAS to ACT Aspire Concordance. 178

References 180

Tables

Table 1.1. Scaling Study—Data Collection Design 6
Table 1.2. Raw Scores of On-Grade Tests for Two Groups Taking Lower-/Upper-Level Scaling Tests. 10
Table 1.3. Raw Scores on Common Items for Two Groups Taking Lower-/Upper-Level Scaling Tests. 11
Table 1.4. Scaling Test Raw Scores, by Grade—English. 18
Table 1.5. Scaling Test Raw Scores, by Grade—Math 19
Table 1.6. Scaling Test Raw Scores, by Grade—Reading. 20
Table 1.7. Scaling Test Raw Scores, by Grade—Science 21
Table 1.8. ACT Aspire English Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade 32
Table 1.9. ACT Aspire Math Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade 32
Table 1.10. ACT Aspire Reading Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade 33
Table 1.11. ACT Aspire Science Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade 33
Table 1.12. ACT Aspire Writing Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade 33
Table 1.13. Lowest Obtainable Scale Scores (LOSS) and Highest Obtainable Scale Scores (HOSS). 34
Table 2.1. ACT Aspire Composite Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade. 36
Table 2.2. ACT Aspire Composite Score Effective Weights Assuming Equal Nominal Weights 37

Table 2.3. ACT Aspire ELA Effective Weights Assuming Equal Nominal Weights 38
Table 2.4. ACT Aspire STEM Effective Weights Assuming Equal Nominal Weights.....39
Table 3.1. 2014 ACT Aspire English Norm Group Demographics 41
Table 3.2. 2014 ACT Aspire Mathematics Norm Group Demographics 43
Table 3.3. 2014 ACT Aspire Reading Norm Group Demographics 45
Table 3.4. 2014 ACT Aspire Science Norm Group Demographics 47
Table 3.5. 2014 ACT Aspire Writing Norm Group Demographics 49
Table 3.6. 2014 ACT Aspire English Norms: Percent of Students at or below Each Scale Score Value 51
Table 3.7. 2014 ACT Aspire Mathematics Norms: Percent of Students at or below Each Scale Score Value 53
Table 3.8. 2014 ACT Aspire Reading Norms: Percent of Students at or below Each Scale Score Value 55
Table 3.9. 2014 ACT Aspire Science Norms: Percent of Students at or below Each Scale Score Value 57
Table 3.10. 2014 ACT Aspire Writing Norms: Percent of Students at or below Each Scale Score Value 59
Table 4.1. Descriptive Statistics of the Sample Used in the Concordance Analysis.....63
Table 4.2. Descriptive Statistics of the ACT Explore and Concorded ACT Aspire Scores in Cross-Sectional Evaluation Sample 66
Table 4.3. Descriptive Statistics of ACT Plan and Concorded ACT Aspire Scores in Cross-Sectional Evaluation Sample.....66
Table 4.4. Descriptive Statistics of ACT Explore Grade 8 and ACT Plan Grade 10 in the Longitudinal Data 75
Table 4.5. Descriptive Statistics of Concorded ACT Aspire in the Longitudinal Data.....75
Table 5.1. ACT College Readiness Benchmark and ACT Readiness Benchmark for Grades 8–10.....77
Table 5.2. Classification Agreement Rate between ACT Aspire and EPAS Benchmarks.....78
Table 5.3. ACT Readiness Benchmarks, Grades 3–7 78
Table 5.4. ACT Readiness Benchmarks in ELA and STEM, Grades 3–10.....79
Table 5.5. Benchmarks, Low Cuts, and High Cuts for ACT Readiness Benchmarks by Subject and Grade 81
Table 6.1. Longitudinal Samples Used to Develop Predicted Paths in 2014.....85
Table 6.2. ACT Aspire Gain Score Means and Standard Deviations 88
Table 7.1. Descriptive Statistics of ACT Composite Scale Scores by Each ACT NCRC Level. .93
Table 7.2. ACT Composite Scale Scores Indicating a 50% Chance of Obtaining Different ACT NCRC Certificates 94
Table 7.3. Descriptive Statistics of the Analysis Sample (*N*=13,528) 94
Table 7.4. EPAS and ACT Aspire Composite Scores Corresponding to ACT NCRC Levels....95
Table 7.5. ACT Aspire Composite Scores Corresponding to the 50% Chance of Obtaining Each ACT NCRC Level.....95

Table 8.1. Raw Score and Scale Score Reliability Coefficient Ranges by Grade for Four ACT Aspire Subject Tests and the Composite Score: Spring 2013 Special Studies Data 98
Table 8.2. Raw Score Reliability Coefficient Ranges by Grade for Four ACT Aspire Tests: Spring 2014 Operational Data 99
Table 8.3. Writing Test Correlations between Rater 1 and Rater 2, by Trait* and by Form: Spring 2013 Special Studies Data 100
Table 8.4. Writing Test Reliability Coefficients Based on Four Trait Scores: Spring 2013 Special Studies Data 100
Table 9.1. Descriptive Statistics for ACT Explore, ACT Plan and ACT Aspire Scale Scores.. 105
Table 9.2. Correlations between ACT Explore/ACT Plan Scale Scores and ACT Aspire Scale Scores..... 106
Table 9.3. Disattenuated Correlations between ACT Explore/ACT Plan Scale Scores and ACT Aspire Scale Scores..... 107
Table 9.4. Scale Score Reliabilities 107
Table 9.5. Multitrait-Multimethod Matrices for ACT Explore/Plan and ACT Aspire Scale Scores by Grade Level 116
Table 9.6. Sample Sizes by Grade and Subject for the Sample of Students with Scores on ARMT+ and ACT Aspire in Spring 2013..... 120
Table 9.7. Descriptive Statistics for ARMT+ and ACT Aspire Scale Scores..... 120
Table 9.8. Correlations between ARMT+ Scale Scores and ACT Aspire Scale Scores 121
Table 9.9. Disattenuated Correlations between ARMT+ Scale Scores and ACT Aspire Scale Scores.....121
Table 9.10. Multitrait-Multimethod Matrices for ARMT+ and ACT Aspire Scale Scores by Grade Level..... 130
Table 9.11. Disattenuated Correlations between ARMT+ and ACT Aspire Scale Scores by Grade Level..... 133
Table 10.1. Percentage of Items Different* Across Online and Paper Forms 137
Table 10.2. Sample Sizes by Grade and Subject..... 138
Table 10.3. English Common-Item Raw Score Summary for Online and Paper Modes..... 139
Table 10.4. Mathematics Common-Item Raw Score Summary for Online and Paper Modes 140
Table 10.5. Reading Common-Item Raw Score Summary for Online and Paper Modes.....141
Table 10.6. Science Common-Item Raw Score Summary for Online and Paper Modes 142
Table 10.7. Writing Common-Item Raw Score Summary for Online and Paper Modes..... 143
Table 10.8. English Scale Score Summary for Online and Paper Modes 156
Table 10.9. Mathematics Scale Score Summary for Online and Paper Modes.....157
Table 10.10. Reading Scale Score Summary for Online and Paper Modes 158
Table 10.11. Science Scale Score Summary for Online and Paper Modes 159
Table 10.12. Writing Scale Score Summary for Online and Paper Modes 160
Table A1. Number of Constructed-response Items and Number of Score Points Included in the On-Grade ACT with Constructed-Response Test Forms in the Scaling Study.....177
Table B1. EPAS to ACT Aspire Concordance 178

Figures

Figure 1.1. Construction of the ACT Aspire scaling tests 4
Figure 1.2. Effect sizes on whole scaling test scores derived from 2PL and 3PL, and effect sizes on individual scaling tests raw scores.. 14
Figure 1.3. Scatter plots of average scale scores against scaling test raw scores, by grade . .22
Figure 1.4. Conditional standard error of measurement, raw scores and scale scores 26
Figure 1.5. Effect sizes computed from the final ACT Aspire scale, whole scaling test scores, and individual scaling test raw scores 30
Figure 4.1. Scatter plots between ACT Aspire scale score and EPAS scale scores. 64
Figure 4.2. Distributions of ACT Explore scale scores and concorded ACT Aspire scale scores 67
Figure 4.3. Distributions of ACT Plan scale scores and concorded ACT Aspire scale scores. .69
Figure 4.4. Box plots of ACT Plan (or concorded ACT Aspire scale scores) conditional on ACT Explore (or concorded ACT Aspire Scores) 71
Figure 6.1. Prototype of ACT Aspire Student Progress Report 84
Figure 6.2. Prototype ACT Aspire Aggregate Progress Report 85
Figure 6.3. ACT Aspire gain score statistics for Grade 3–4 Mathematics 89
Figure 7.1. Sample Progress toward Career Readiness indicator from ACT Aspire report. 92
Figure 9.1. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire English scale score. 108
Figure 9.2. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Mathematics scale score 109
Figure 9.3. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Reading scale score. 111
Figure 9.4. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Science scale score 112
Figure 9.5. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Composite scale score 114
Figure 9.6. Box plots of ARMT+ mathematics scale scores for each ACT Aspire Mathematics scale score 122
Figure 9.7. Box plots of ARMT+ reading scale scores for each ACT Aspire Reading scale score 125
Figure 9.8. Box plots of ARMT+ science scale scores for each ACT Aspire Science scale score 128
Figure 10.1. Plots of cumulative percent of students for common item raw scores across mode 144
Figure 10.2. Plots of cumulative percent of students for scale scores across mode 161
Figure 10.3. Box plots of scale scores by grade and mode for each subject area 171

Overview

ACT Aspire® includes a vertically scaled battery of achievement tests designed to measure student growth in a longitudinal assessment system for grades 3–10 in English, reading, writing, mathematics, and science. ACT Aspire is designed to measure students' progress toward college and career readiness. The scale scores are linked to college and career data through scores on The ACT® and the ACT National Career Readiness Certificate™ (ACT NCRC®) program. Empirically based ACT College Readiness Benchmarks provide information about whether students are on target for readiness at the appropriate grade/subject levels. To enhance score interpretation, reporting categories for ACT Aspire use the same terminology as the ACT College and Career Readiness Standards (ACT CCRS) and other standards that target college and career readiness (including the standards of many states and the Common Core State Standards [CCSS]). Some reporting categories are unique to ACT Aspire. These include science, technology, engineering, and mathematics (STEM), justification and explanation in mathematics, progress with text complexity in reading, and a progress toward career readiness indicator.

The types of items based on a given construct are determined by considering the amount and nature of the evidence needed to support an inference. These requirements are balanced with maintaining manageable administration conditions. The ACT Aspire design includes several item types (i.e., selected-response, constructed-response, technology-enhanced) and a range of item difficulties at varying depths of knowledge. ACT Aspire assessments cover learning progressions from foundational concepts to sophisticated applications.

Taken as individual subject tests or as a battery, ACT Aspire can be delivered online or as a paper administration.

Introduction

This bulletin provides technical details regarding ACT Aspire™ scale scores, including the development of the scales, description of scores, development of the ACT Readiness Benchmarks, development of a Progress toward Career Readiness indicator, description of norms, and description of equating. This bulletin also summarizes evidence regarding ACT Aspire reliability, validity, and mode comparability. For details about ACT Aspire test development and content, see *ACT Aspire Summative Assessment Technical Bulletin #1* (ACT 2014b).

Analyses described below primarily include data from three large-scale studies conducted during the spring of 2013. Students included a national sample in grades 3–11. ACT Aspire test forms included English, Mathematics, Reading, Science, and Writing.[1] The three studies included (a) a scaling study to develop the ACT Aspire scale, (b) an equating study to place multiple ACT Aspire test forms onto the ACT Aspire scale, and (c) a mode comparability study to link paper and online forms and to evaluate potential mode effects across forms. Details regarding each of these studies are provided in separate sections of this bulletin.

Data from the three studies were also combined with other data sources for particular sections of this bulletin. For example, ACT Aspire scores were matched for students taking other academic achievement tests to evaluate the validity of the interpretation of ACT Aspire scores as measures of academic achievement.

This document should be viewed as an abbreviated technical manual, providing information about the scaling, mode comparability, and other technical characteristics of the ACT Aspire assessments. More complete documentation, including additional technical analyses and additional results from the inaugural operational administration of ACT Aspire in spring of 2014, is forthcoming.

[1] An ACT form was also included for grade 11 for scaling and is described below.

CHAPTER 1

ACT Aspire Score Scale

ACT Aspire Scaling Philosophy

ACT ascribes to a domain definition of growth, where student achievement over the entire range of content is considered (Kolen and Brennan 2014). This conceptualization of what growth means leads naturally to the scaling test data collection design. The scaling test design involves creation of a single test that covers the range of content and difficulty across the domain. This test is then administered to students across the full range of grade levels covered and vertical scaling is used to place performance of students across grade levels onto the same scale. The next section describes in detail the scaling study used to establish the ACT Aspire score scales.

Scaling Study

Introduction to the Scaling Study Design

The purpose of the scaling study is to establish a vertical scale for the ACT Aspire English, Mathematics, Reading, and Science assessments. The ACT Aspire assessments include seven grade level tests per subject: grades 3–8 and early high school (EHS), which is given to students in grades 9 and 10. In addition, by including grade 11 in the scaling study, the vertical scale could be extended to include The ACT (see Appendix A), even though The ACT is not included in the ACT Aspire assessments.

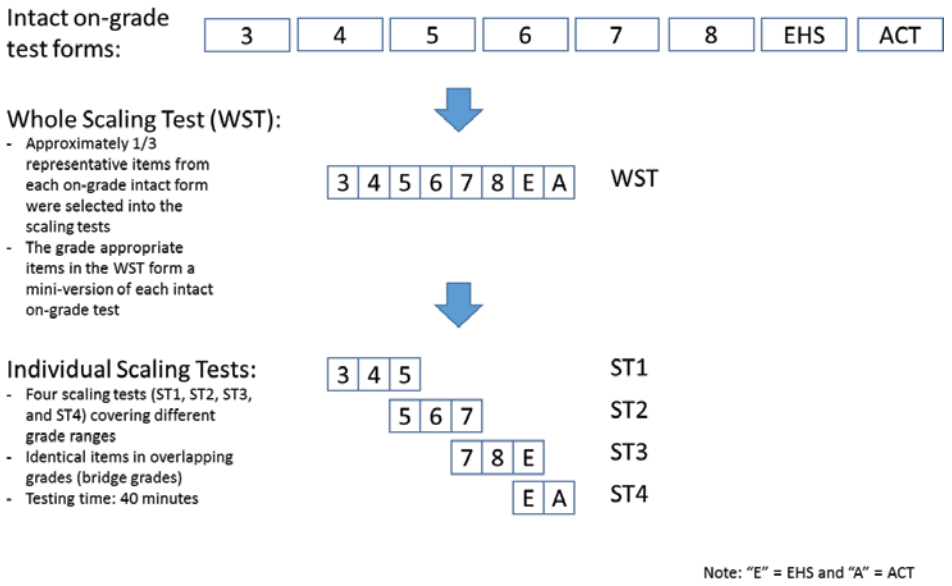
The scaling test design was adopted to create the vertical scale for each subject. Under this design, each student completes two tests: an on-grade test appropriate for his/her grade level and a scaling test with items from multiple grades. The vertical scale is defined using the scaling test, and the scores for each on-grade test are linked to the scale through the scaling test.

Construction of Scaling Tests

Design of the scaling tests is a key component in creating a sound vertical scale under the scaling test design. Items from the scaling tests should cover all content that defines the vertical scale and be sensitive enough to measure growth. Since the vertical scale covers grade 3 to grade 11, it is ideal to construct one scaling test per subject with items from all grade levels and administer this single scaling test to students from grades 3 to 11.

However, to include a sufficient number of items from each content domain in the scaling test, this single test consisting of eight levels of tests (i.e., grades 3–8, 9/10, and 11) would be much too long to administer operationally and would require students in certain grades to complete items which are significantly above or below their grade level and likely achievement level. The length of the scaling test and the prospect of, for example, administering high school students grade 3 items and third graders items intended for high school students, would be unlikely to result in good measurement. To create scaling tests of reasonable length, a single scaling test covering items from grades 3 to 11 (hereafter referred to as the whole scaling test and abbreviated WST) was broken into four separate tests in this study, each of which includes items from two or three consecutive grades with one grade of overlap (referred to as bridge grades) between tests. Specifically, scaling test 1 (ST1) includes items from grades 3 to 5, scaling test 2 (ST2) includes items from grades 5 to 7, scaling test 3 (ST3) includes items from grades 7 to EHS, and scaling test 4 (ST4) includes items from EHS and The ACT (see figure 1.1).

Figure 1.1. Construction of the ACT Aspire scaling tests



The scaling tests were designed to represent a miniature version of each on-grade test. The items chosen for the miniature version of each on-grade test were selected to be representative with respect to both content coverage and item statistics to the extent possible (approximately one-third of the items on an on-grade form for each grade were used in the scaling tests). There were no common items between the scaling test and the on-grade test administered to a student.

Within each scaling test, items were grouped and sorted by grade levels so that lower-grade items were positioned earlier in the test than upper-grade items. Items from the bridge grades (5, 7, and 9/10) in the four scaling tests were identical. The only difference was that these items appear at the end of the lower-level scaling test but at the beginning of the upper-level scaling test. The time limit for each scaling test was 40 minutes.

Data Collection Design

Each student participating in the scaling study was asked to take on-grade tests at the appropriate grade level in all five subjects (English, math, reading, science, and writing) and two scaling tests of different subjects. Once data were collected, analyses to create the scale were conducted separately for each subject. Students from grades 9 and 10 were combined for analysis of the EHS test.[2] Each test could be administered in a separate sitting, and students took the on-grade tests before taking any scaling tests. All testing occurred between April and May in 2013, and all assessments in the scaling study were given online. In total, more than 37,000 students from over 20 school districts and 14 states participated in the 2013 scaling study.

Table 1.1 presents the data structure used to analyze each subject. Each row lists the student grade level and the columns list the on-grade or scaling test assigned to students. Bridge grades are listed twice, one row as "A" for those students assigned to a lower-grade scaling test and one row as "B" for those assigned the upper-grade scaling test. The second column lists the on-grade test form taken by students. The third column lists the scaling test taken by students. For example, grade 7 students (bridge grade) assigned to the "A" group would take the grade 7 on-grade test and the grades 5–7 scaling test (ST2). Because grades 9 and 10 were bridge grades, grade 9 students completed the EHS on-grade test plus either the lower-grade scaling test (ST3) or the upper-grade scaling test (ST4). Grade 10 followed a similar pattern: grade 10 students completed the EHS on-grade test plus ST3 or ST4.

[2] While grade 9 and 10 were combined for analysis of the EHS test, norms, growth, benchmarks, and other information reported to students are separately obtained for grades 9 and 10 within the EHS test. Details are provided in other chapters of this bulletin.

Table 1.1. Scaling Study—Data Collection Design

Student Grade	On-Grade Test									Scaling Test			
3	3									ST1			
4		4								ST1			
5A			5							ST1			
5B			5								ST2		
6				6							ST2		
7A					7						ST2		
7B					7							ST3	
8						8						ST3	
9A							EHS					ST3	
9B							EHS						ST4
10A							EHS					ST3	
10B							EHS						ST4
11								ACT					ST4
12								ACT					ST4

Note: ST1 includes items from grades 3–5; ST2 incudes items from grades 5–7; ST3 includes items from grades 7–EHS; ST4 includes items from EHS and The ACT.

In the scaling study, each student in each grade completed one scaling test and one on-grade test of the same subject. Approximately twice as many students were sampled for bridge grades (5, 7, and 9/10) compared to those from non-bridge grades to allow half of the bridge-grade students to take the lower-level scaling test and half to take the upper-level one. All students in the same grade took the same on-grade test. The two groups taking lower-/upper-level scaling tests in each bridge grade were designed to be randomly equivalent. Once lists of recruited students were available, students at a bridge grade were randomly assigned to take the upper- or lower-level scaling test by spiraling tests across students. For example, for four students at a bridge grade within the same classroom, if the first student was assigned the lower-level scaling test, the second would be assigned the upper, the third the lower, and the fourth the upper. This spiraling pattern continued for all students in this classroom and continued into other classrooms and schools at the same grade level.

Creating the Vertical Scale

Under the scaling test design, the vertical alignment of student performance across all grades was obtained from the scaling tests. Once the alignment was established, a scale with desirable properties (i.e., target mean, standard deviation, standard error

of measurement, number of scale score points, etc.) was created. Then, the base form of each on-grade test was linked to the vertical scale. As new on-grade forms are developed, they are horizontally equated to the base form to maintain the vertical scale (see chapter 11).

The process of establishing the ACT Aspire vertical scale can be summarized in three steps:

1. Link across the four scaling tests to establish the vertical relationship from grades 3 to 11 in a subject.
2. Create the scale with desired properties based on the linked scaling test.
3. Link each on-grade test to the vertical scale.

Detailed descriptions of these three steps are presented below.

Step 1: Link across the Four Scaling Tests

The goal of this step is to link the four separate scaling tests so that scores of students taking different scaling tests were put on the same scale. Note that if the whole scaling test were given to all students, this step would not be necessary. To conduct the linking, ST3 (composed of items from grades 7 to EHS) was selected as the base test. This decision was made because (1) it contains items at the top of the ACT Aspire scale, (2) two of the three other scaling tests (ST2 and ST4) can be linked to this base test directly, and (3) it is adjacent to ST4, which was used to link The ACT with constructed-response tests (see appendix A) to the ACT Aspire scale.

There are multiple possible ways of placing the four scaling tests on the same scale. The options can be broken into two dimensions: linking design options and statistical method options.

Linking Design Options

There were three linking design options:

1. Link through random equivalent groups in the bridge grades using students who took the upper- or lower-level scaling tests (e.g., grade 5 students taking ST1: grades 3–5 or ST2: grades 5–7).
2. Link through common items between adjacent scaling tests (e.g., common grade 5 items in ST1 and ST2).
3. Treat the entire on-grade test as an external anchor item set taken by students in the bridge grades.

Each of these design options makes certain assumptions which should be met. The first design assumes the two groups are randomly equivalent. The second design assumes common items perform similarly between the two scaling tests. The third design assumes that the common items on the same on-grade test form perform similarly across similar groups of students.

Statistical Method Options

Three methodology options were considered for scaling, with variations under each: (1) the Thurstone method, (2) an ad hoc method based on linear or equipercentile linking of different scaling tests (hereafter referred to as the *ad hoc method*), and (3) the IRT method.

The Thurstone method (1938) assumes that normalized raw scores (*z*-scores) are normally distributed within each grade (which also implies that *z*-scores across grades are linearly related). Depending on the data collection design used, means and standard deviations of a selected range of *z*-scores of students from different grades on a common set of items or *z*-scores from equivalent groups of students on different sets of items can be used to establish the relationships between standardized scale scores across grades. Then, based on the relationship obtained, by fixing the mean and standard deviation of one grade to the desired scale score moments or fixing the means of any two grades, the means and standard deviations of scale scores of all grades are obtained. To obtain the final raw-to-scale conversion, the normalized raw scores of each grade are linearly transformed to match the scale score mean and standard deviation obtained earlier for that grade.[3]

Besides the potential differences in results based on different data collection designs, Thurstone method results can also vary depending on the range of score points selected for conducting the analyses. Gulliksen (1950, 284) suggested selecting ten or twenty score points when applying the Thurstone method. Williams, Pommerich, and Thissen (1998) explored two versions of score-point selection: one with all the score points observed for both grades, and one with score points between 10% and 90% of the distributions of both grades. This study used all score points as the primary choice, but also experimented with a few other range selections: 2.5%–97.5%, 5%–95%, and 10%–90%. The purpose of trying various trimming options was to gauge the effects on the scale obtained from the Thurstone method.

The ad hoc method involves linking any possible pair of scaling tests using either the linear or the equipercentile method (see Kolen and Brennan 2014). The goal of this linking was to predict students' total raw scores on the whole scaling test based on the individual scaling test they had taken, and then using the predicted whole scaling test raw scores as a basis for constructing the vertical scale. Specifically, to estimate the whole scaling test score for a student who only took one individual scaling test, we first estimate the student's predicted scores on the other three scaling tests and then sum the four individual scaling test raw scores to obtain the whole scaling test score. To predict scores on the three scaling tests that the student did not take, numerous pair-wise linkages were conducted. For example, to predict ST2, ST3, and ST4 for students who took ST1, the following links are needed: ST1 to ST2, ST1 to ST3 through ST2, ST1 to ST4 through ST2 and ST3. There were different options of computing the whole scaling test score depending on how common item scores

[3] See chapter 9 in Kolen and Brennan (2014) for more details about this method.

were treated. One option was to sum up the raw and estimated raw scores on all four scaling tests, in which case common items were double counted. Other options included different ways of de-counting the common items score, such as using the simple sum score minus the common item scores from one of the scaling tests or minus the average of scores on the two sets of common items. Whenever common item scores were required in computing the whole scaling test scores, additional links were needed among the common items from different scaling tests to account for the common item total scores.

The item response theory (IRT) linking method involves putting IRT item parameters from the four scaling tests on the same scale. Both the three-parameter logistic (3PL) model and the two-parameter logistic (2PL) model were considered for the dichotomously-scored items. The generalized partial credit model (GPC) was used for polytomously-scored (i.e., constructed-response) items. The 3PL contains three parameters for an item plus a parameter (theta) for student proficiency, and the 2PL contains two parameters for an item plus a parameter for student proficiency. The GPC is analogous to the 2PL but incorporates more than two score categories.[4] All calibrations were conducted using single group analysis with PARSCALE software. Each of the four scaling tests was first independently calibrated. Under the common item design, Stocking-Lord method (Stocking and Lord 1983) was used to transform nonbase scaling test item parameters to the base test (ST3); under the random equivalent groups design, students' mean and standard deviation (SD) of theta scores between the two groups in the bridge grade were used to compute the transformation constants (see Kolen and Brennan 2014, 180–182). Students with extreme theta scores (less than or equal to −6 or greater than or equal to +6) were excluded from the calculation of mean and SD, because these theta scores were fixed by arbitrary boundaries set up in the scoring software. The obtained scale transformation slope and intercept were then applied to item parameter estimates from each independent calibration to generate the scaled item parameters for all scaling tests. Students' theta scores were then estimated from the scaled item parameters which enabled them to be put on the same scale.

Design and Method Selection

In theory, any combination of the design options and statistical methods could be used. During the exploratory stage of analysis, many combinations were considered. To limit the options, preliminary evaluations of results from different linking methods were conducted. Four primary issues were considered: (1) determining whether the required assumptions of each method were met, (2) application of results from a simulation study to facilitate decision making, (3) checking to see whether growth patterns (i.e., effect size between adjacent grades) among different results were consistent or different, and (4) verifying that the desired scale properties could be realized under a specific method.

[4] For further descriptions of IRT and the 2PL, 3PL, and GPC models see Baker and Kim (2004), de Ayala (2009), or Yen and Fitzpatrick (2006).

Design selection. The data collection design used to develop the vertical scale was determined primarily after checking whether certain assumptions were met.

The random equivalent groups design linked adjacent scaling tests through the two groups taking lower-/upper-level scaling tests in each bridge grade by assuming the two groups were equivalent. Since spiraling occurred at the student level within each grade in a school, ideally, the number of students taking the lower-level scaling test should be very close to that of students taking the upper-level test in each grade of each school. During the data cleaning, if all students in the bridge grade in one school only took one level of a scaling test, or if the ratio of students taking one level test over the other level was more than 2, all students in that grade from that school were removed from the final analysis. This was done to ensure group equivalence. To further evaluate whether the two groups were equivalent after data cleaning, on-grade test raw scores were compared across groups because they took the same on-grade test form. Table 1.2 presents sample size, mean, and standard deviation of on-grade test scores for the two groups in the bridge grades. For each pair of groups, an independent-samples *t*-test was conducted to evaluate the statistical significance of test raw score means (at 0.05 significance). None of the *t*-tests were statistically significant—an indication that scores for the two groups from each bridge grade did not differ.

The common item design involved linking through the common items between adjacent scaling tests and assumed common items performed similarly between the two scaling tests. However, context effects were a concern under the common item

Table 1.2. Raw Scores of On-Grade Tests for Two Groups Taking Lower-/Upper-Level Scaling Tests

		Group A			Group B		
Subject	**Grade**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
English	5	845	14.588	4.345	871	14.701	4.456
	7	889	19.580	5.898	928	19.496	5.820
	9/10	667	26.187	10.461	687	25.128	10.431
Math	5	868	10.058	3.382	870	10.176	3.459
	7	863	15.651	6.453	882	15.593	6.239
	9/10	688	17.247	8.440	706	17.479	8.927
Reading	5	822	15.798	5.492	833	15.739	5.717
	7	806	13.166	5.241	800	13.340	5.364
	9/10	707	14.651	7.058	684	14.415	7.088
Science	5	727	18.171	7.132	721	18.337	7.104
	7	819	18.812	7.511	834	18.801	7.955
	9/10	634	14.218	8.082	648	14.520	7.772

Note: Group A took the lower level ST, and Group B took the upper level ST.

design because common items appeared at the beginning of the upper-level scaling test and at the end of the lower-level scaling test. Specifically, the concern was that the items might appear easier when they appeared at the beginning of the scaling test than at the end of the test.

Context effects were examined by comparing the total raw score points on the common items between the two groups of students taking lower/upper-level scaling tests. Since it was previously determined that the two groups at the bridge grades, one taking the upper-grade scaling test (denoted as B) and one taking the lower-grade scaling test (denoted as A), were randomly equivalent, the average raw scores on the common item set across groups should be similar. However, if context effects are present, non-ignorable mean differences on the common items across groups might be expected. For example, it has been shown that the groups taking scaling test 5A (grade 5, lower-grade scaling test) and 5B (grade 5, upper-grade scaling test) are equivalent, so if the average raw score on grade 5 items in the scaling test for 5A is different from that for 5B, context effects likely exist. Table 1.3 shows descriptive statistics of lower- and upper-scaling test raw scores for groups taking common items. Group B scored higher on the common items for all subjects and all bridge grades. Items appearing at the beginning of the test (group B) appeared to be easier than when the same items appeared at the end of a scaling test (group A). The *t*-tests were statistically significant (at the 0.05 level) for grades 5 and 7 in all subjects and grade 9/10 in math and science. Context effects appeared to affect performance on the common items, which is a violation of the assumption that items should perform similarly under the common-item design. Therefore, this design was not adopted.

Table 1.3. Raw Scores on Common Items for Two Groups Taking Lower-/Upper-Level Scaling Tests

		Group A Taking Lower-Level ST			Group B Taking Upper-Level ST		
Subject	**Grade**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**
English	5*	845	10.174	3.818	871	11.200	3.292
	7*	889	9.529	3.907	928	10.335	3.534
	9/10	667	12.337	5.614	687	12.582	5.280
Math	5*	868	4.358	1.930	870	4.670	1.952
	7*	863	3.928	2.299	882	4.195	2.296
	9/10*	688	3.911	2.954	706	4.249	2.840
Reading	5*	822	3.456	1.720	833	3.619	1.635
	7*	806	4.227	2.262	800	4.591	2.309
	9/10	707	4.494	3.225	684	4.722	3.182
Science	5*	727	3.884	2.129	721	4.277	2.080
	7*	819	3.891	1.984	834	4.125	1.910
	9/10*	634	6.412	4.302	648	7.110	4.180

**t*-test $p < .05$

Under the common on-grade linking design, where the same test form was administered to two similar groups of students, we assumed that the common items performed similarly across groups. Results from this design were compared with those derived from the random equivalent groups design. Specifically, both designs were tried out using the Thurstone and ad hoc methods. Results indicated that the growth patterns resulting from these two designs were very similar to each other. This was not surprising, since the two groups were shown to be equivalent and had comparable average raw scores on the on-grade test. Because the score patterns appeared similar across common on-grade linking and random equivalent groups linking and, more importantly, because the random groups assumption was tenable, the four scaling tests were linked using the random equivalent groups design.

Method selection. Thurstone, ad hoc, and IRT statistical methods were tried using the random equivalent groups linking design. Of the two non-IRT methods (i.e., the Thurstone and ad hoc linking methods), the growth patterns in terms of effect size between adjacent grades were similar. Each of the non-IRT methods had problematic disadvantages. The ad hoc method involved many linkages and may introduce unnecessary linking errors. A disadvantage of Thurstone scaling is that because it does not involve recovery of the whole scaling test raw scores, the conditional standard error of measurement (CSEM) on the whole scaling test cannot be readily evaluated or stabilized, though methods are available to obtain constant CSEM for the on-grade tests. CSEM properties, described below, were an important aspect of the development of the ACT Aspire scale.

Under the IRT method, 2PL and 3PL models were fit to the data for dichotomous items. The final model choice was informed by literature reviews, previous internal studies examining the performance of the 2PL and 3PL using both simulated data and real data, and the comparison of growth patterns between the two models on the scaling tests.

Compared with the 2PL model, the 3PL model is more widely used in other large-scale tests. Due to the inclusion of an additional item parameter, the 3PL model typically fits the data better. In addition, the fact that the 3PL includes a parameter intended to account for guessing often makes it a defensible model for multiple-choice items. However, based on studies conducted internally, it was found that the 3PL had some practical limitations, which have been mentioned elsewhere (Yen and Fitzpatrick 2006).

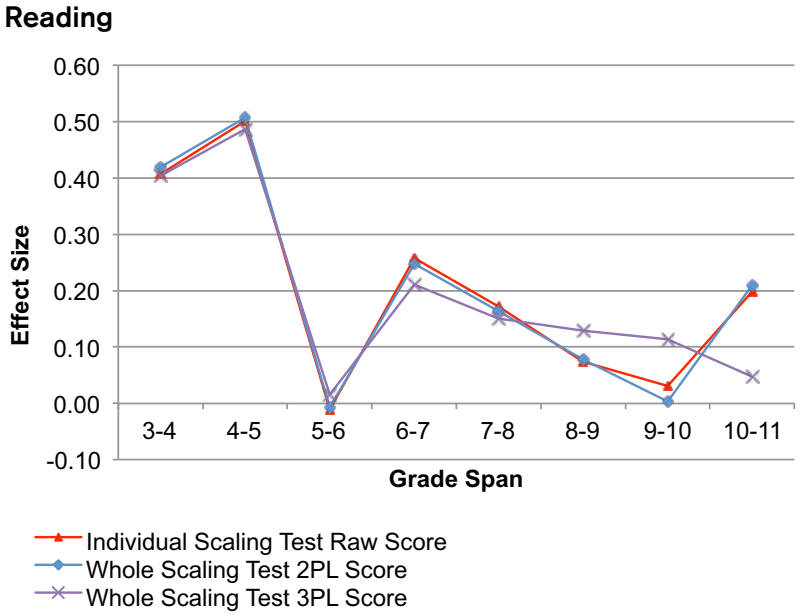
One major issue with the 3PL was related to *c*-parameter estimation (commonly referred to as the pseudo-guessing parameter). For example, the 3PL estimation more frequently led to a less-than-optimal solution (or none at all), which resulted in unstable item parameter estimates (or no estimates). In such cases, different IRT calibration software programs produced different results (e.g., PARSCALE versus BILOG-MG). These problems appeared to be due to estimation of the c-parameter, which also impacted both *a*- and *b*-parameter estimates (slope and difficulty parameters, respectively). On the other hand, the 2PL is a more parsimonious model than the 3PL and had stable item parameter estimates when different calibration samples and different software were used. An even more parsimonious model is the Rasch model, but preliminary results indicated that the presence of a slope parameter that varies across items, which is included in the 2PL and 3PL, was advantageous for creating the ACT Aspire vertical scale.

Another potential benefit of using the 2PL compared to the 3PL is that it can often be used to obtain reliable parameter estimates from smaller sample sizes. In the 2013 scaling study, we had adequate sample sizes for most grades, but the sample sizes for some tests were relatively low for conducting some IRT calibrations (e.g., between 1,000 and 1,500).

To evaluate the effects of the IRT model on the patterns of scores across grade-level growth patterns in terms of effect size on the predicted whole scaling test, scores[5] were compared between results from the 2PL and 3PL. Effect size was computed as the difference between the mean scores of adjacent grades divided by the square root of the average variances for the two groups (Kolen and Brennan 2014, 461). Using the Reading test as an example, the grade-to-grade growth in effect size derived from the whole scaling test scores using 2PL and 3PL are plotted in figure 1.2. Also included in the plot are the effect sizes computed from the individual scaling test raw scores. For example, the grade 3 to grade 4 effect size was computed using ST1 raw scores on grades 3 and 4 students. For the bridge grades, the effect size was computed between one of the random groups and its adjacent grade taking the same scaling test. For example, effect size between grades 4 and 5 was computed on ST1 raw scores between grade 4 and 5A students; effect size between grades 5 and 6 was computed on ST2 raw scores between grade 5B and 6 students. Since 5A and 5B groups were shown to be equivalent, either group could be treated as representative of the entire group. The effect size between grades 9 and 10 was the average of effect sizes for 9A versus 10A and 9B versus 10B.

[5] These were true score estimates obtained by applying the scaled item parameter estimates of all items on the whole scaling test given scaled proficiency (theta) estimates of each student obtained from the part of the whole scaling test each student was actually administered.

Figure 1.2. Effect sizes on whole scaling test scores derived from 2PL and 3PL, and effect sizes on individual scaling tests raw scores.



The purpose of comparing whole scaling test effect sizes and individual scaling test raw score effect sizes was to ensure the process of linking the four scaling tests did not distort the grade-to-grade relationships in the scaling tests. As shown in figure 1.2, the effect size pattern between 3PL and 2PL is very similar for grades 3–8 but diverges from grades 8 to 11. The 2PL results are closer to that of the raw scaling test scores.

In summary, after checking assumptions and referring to simulation results, the random equivalent group linking design and IRT 2PL model were selected as the method to link across the four scaling tests.

Step 2: Create the Vertical Scale Based on the Linked Scaling Tests

Generate the Projected Whole Scaling Test Raw Scores

After step 1 was completed, IRT item parameters for all four scaling tests were placed on the same scale. Students' theta scores were also on the same scale when estimated from the scaled item parameters. Based on the scaled item parameter estimates and theta estimates, students' true scores on the whole scaling test, which were on the raw score metric, were estimated from the IRT test characteristic curve (TCC) of the whole scaling test.

Create an Interim Scale with Constant Conditional Standard Error of Measurement (CSEM)

The conditional standard error of measurement (CSEM) of the raw scores on a test is typically an inverted U shape with the CSEM values being much smaller at the two ends. To stabilize the CSEM along the score scale, raw scores on the whole scaling test were nonlinearly transformed to an interim score scale using the arcsine transformation (e.g., see Kolen and Brennan 2014, 405). Although this stabilization was applied to the scaling test to obtain a constant CSEM property for the ACT Aspire scale, after the on-grade tests were linked to the whole scaling test, there was no guarantee that the constant CSEM property would be maintained on the on-grade tests. Empirical results indicate that applying the arcsine transformation did help to stabilize the scale score CSEM of the on-grade test after it was linked to the whole scaling test.

To compute the CSEM, an extended Lord-Wingersky recursive algorithm (Hanson 1994; Wang, Kolen, and Harris 2000, 219; Kolen and Brennan 2014, 199) was adopted to obtain the expected distribution on the interim scale score for each student based on his/her theta score. The standard deviation of the expected score distribution was the CSEM of the interim score scale for that student. Weights were used to equalize the contribution of students from different grade levels to the CSEM calculation because sample sizes varied across grades, particularly for the bridge grades where the data collection design resulted in larger samples of students compared to other grades (e.g., see sample sizes listed in tables 1.4–1.7).

Linear Transformation of the Interim Scale to the Final Scale

The interim scale scores, with the property of constant CSEM, were then linearly transformed to the ACT Aspire scale. Since linear transformation does not change the relative magnitude of the CSEMs, the constant CSEM property was carried over to the final scale. The linear transformation was selected so that the standard error of measurement on the whole scaling test was about 1.8 scale score points and the mean was set to an arbitrary value to anchor the scale.

Step 3: Link the On-Grade Test to the Vertical Scale

The last step in the vertical linking process was to link each of the on-grade English, Mathematics, Reading, and Science tests to the vertical scale established using the whole scaling test in each subject. A single group linking design was adopted, since the same group of students took both the on-grade test and the scaling test in the same subject.

The IRT observed score linking method (Kolen and Brennan 2014, ch. 6) was used to create the link. First, each on-grade test form was calibrated independently. Second, item parameter estimates from the on-grade test were transformed to the whole scaling test scale by matching the mean and standard deviation of theta scores estimated from each on-grade test to theta scores estimated from the scaled item parameters of the scaling tests. Third, the estimated distribution of number-correct scores (using scaled item parameters) on the on-grade form was linked to that of the whole scaling test. The IRT observed score linking yields the raw-to-scale conversion for each on-grade test. The scale scores were rounded, truncated, and shifted to desired score ranges to provide the final reportable scale scores.

Evaluation of the Scales

Once the scale was created and students' scale scores were generated, the following analyses were conducted to evaluate the scales:

1. *Check whether the scale maintains the on-the-same-scale property.* When multiple grades of students took the same scaling test, their scores on that scaling test were automatically on the same scale. The entire scaling process for ACT Aspire involved multiple steps, including connecting individual scaling tests and linking the on-grade test to the whole scaling test. Therefore, it was important to verify that ACT Aspire scale scores obtained from on-grade tests could maintain the vertical relationships obtained from the scaling test. This property is referred to as the on-the-same-scale property. To evaluate this property, scaling test raw score on each of the four scaling tests were used as a reference to define the observed relationships across grade level. If the *on-the-same-scale* property was maintained, students who had the same raw scores on the scaling test should have similar scale scores on the on-grade test, even though their scale scores are from on-grade tests at different grade levels.
2. *Evaluate the constant CSEM property.* As mentioned above, the CSEM stabilization process was applied to the scale defined by the scaling test. To evaluate whether this property was maintained for the on-grade tests, CSEM curves for each on-grade test were plotted and examined.
3. *Evaluate the growth pattern.* The growth pattern derived from ACT Aspire scale scores were examined and compared against the growth pattern derived from the raw scaling test scores.

Results

Scaling Test Raw Scores

Tables 1.4–1.7 present the mean and standard deviations of raw scores on individual scaling tests and the mean and standard deviations of raw scores on the portion of grade-specific items in each test. For example, the grade 3 students' average raw score is 22.525 for the English scaling test (ST1) and when items are grouped by grade levels, the average score is 7.313 on all grade 3 items, 7.481 on all grade 4 items, and 7.731 on all grade 5 items. The average scores on items from different grade levels (i.e., across the columns) are not directly comparable since the *number of items is different across grade levels*. Similarly, average scaling test raw scores are only comparable among groups who took the same scaling test since the numbers of score points and most of the items are different across the four scaling tests.

Within each scaling test, summarizing the average total raw score on the scaling test by item grade-level provides diagnostic information regarding where growth occurs. For example, a slightly negative growth occurs between grade 5 students (05B) and grade 6 students on ST2 in reading. Reviewing the means on items grouped by grade levels reveals that the reverse growth occurred on the grade 6 items. Another example of negative growth is on ST4 in reading between group 9B and 10B. Group 10B performed worse on both the EHS and The ACT items.

Context effects are also observable in tables 1.4–1.7 between the bridge-grade groups of students ("A" and "B" groups), because the common item mean scores are always higher in the B group (where common items are given at the beginning of the test) than the A group (where common items are given at the end of the test). This is true for all subjects and all bridge grades. As explained earlier, this was one reason why the common item linking design was not used to develop the vertical scale.

The standard deviations of raw scores tended to increase as grade level increased within the same scaling test. In other words, group variability increased as students advance in grade level.

Evaluating the On-the-Same-Scale Property

Figure 1.3 presents plots of the average scale scores per grade against the scaling test raw score, by scaling test. As can be seen, students with the same scaling test raw score have similar scale scores, on average, regardless of which on-grade tests they have taken. These results are consistent with an interpretation of scores of on-grade tests as being on the same vertical scale.

Table 1.4. Scaling Test Raw Scores, by Grade—English

Scaling Test	Grade	N	Mean (SD) on ST	Mean (SD) of Grade Specific Items in Scaling Tests							
				G3	G4	G5	G6	G7	G8	G_EHS	G_ACT
ST1	3	1810	22.525 (7.975)	7.313 (2.542)	7.481 (3.308)	7.731 (3.499)					
	4	1617	26.078 (8.320)	8.208 (2.585)	8.730 (3.347)	9.140 (3.689)					
	05A	845	29.060 (8.631)	9.058 (2.713)	9.828 (3.380)	10.174 (3.818)					
ST2	05B	871	28.021 (8.704)			11.200 (3.292)	8.503 (2.998)	8.318 (3.582)			
	6	1824	28.548 (9.528)			11.386 (3.604)	8.490 (3.189)	8.672 (3.834)			
	07A	889	30.670 (9.678)			12.048 (3.496)	9.093 (3.280)	9.529 (3.907)			
ST3	07B	928	29.763 (10.335)					10.335 (3.534)	8.672 (3.524)	10.755 (4.514)	
	8	1436	31.769 (11.059)					10.818 (3.683)	9.171 (3.677)	11.781 (4.807)	
	09A	343	32.402 (12.237)					11.117 (3.975)	9.172 (4.042)	12.114 (5.228)	
	10A	324	32.710 (13.787)					10.843 (4.267)	9.293 (4.410)	12.574 (5.995)	
ST4	09B	368	29.481 (12.552)							12.332 (5.122)	17.149 (8.281)
	10B	319	31.367 (14.036)							12.871 (5.450)	18.495 (9.288)
	11	760	32.925 (13.797)							13.943 (5.206)	18.982 (9.356)

Table 1.5. Scaling Test Raw Scores, by Grade—Math

Scaling Test	Grade	N	Mean (SD) on ST	Mean (SD) of Grade Specific Items in Scaling Tests							
				G3	G4	G5	G6	G7	G8	G_EHS	G_ACT
ST1	3	1793	9.090 (3.573)	3.732 (1.944)	2.697 (1.472)	2.661 (1.503)					
	4	1624	11.580 (4.360)	4.573 (2.029)	3.345 (1.841)	3.662 (1.743)					
	05A	868	13.778 (5.169)	5.315 (2.133)	4.105 (2.212)	4.358 (1.930)					
ST2	05B	870	9.924 (3.434)			4.670 (1.952)	2.600 (1.237)	2.654 (1.501)			
	6	1751	10.817 (4.216)			4.812 (2.047)	2.939 (1.428)	3.065 (1.752)			
	07A	863	12.818 (5.17)			5.395 (2.129)	3.495 (1.693)	3.928 (2.299)			
ST3	07B	882	9.347 (4.689)					4.195 (2.296)	2.407 (1.508)	2.745 (1.942)	
	8	1451	10.580 (5.289)					4.504 (2.369)	2.682 (1.505)	3.394 (2.375)	
	09A	381	11.213 (6.218)					4.559 (2.704)	2.877 (1.679)	3.777 (2.719)	
	10A	307	12.160 (7.023)					5.156 (2.836)	2.925 (1.745)	4.078 (3.219)	
ST4	09B	376	10.580 (5.691)							4.133 (2.632)	6.447 (3.427)
	10B	330	11.361 (6.926)							4.382 (3.060)	6.979 (4.269)
	11	669	12.765 (6.632)							5.139 (3.041)	7.626 (4.087)

Table 1.6. Scaling Test Raw Scores, by Grade—Reading

Scaling Test	Grade	N	Mean (SD) on ST	Mean (SD) of Grade Specific Items in Scaling Tests							
				G3	G4	G5	G6	G7	G8	G_EHS	G_ACT
ST1	3	1769	8.077 (4.054)	3.003 (1.686)	2.757 (1.879)	2.317 (1.451)					
	4	1569	9.811 (4.467)	3.516 (1.755)	3.587 (2.040)	2.709 (1.573)					
	05A	822	12.102 (4.67)	4.153 (1.710)	4.493 (2.154)	3.456 (1.720)					
ST2	05B	833	12.552 (4.903)			3.619 (1.635)	5.261 (2.322)	3.672 (2.036)			
	6	1759	12.495 (5.357)			3.624 (1.695)	5.077 (2.422)	3.794 (2.179)			
	07A	806	13.878 (5.384)			3.970 (1.685)	5.681 (2.388)	4.227 (2.262)			
ST3	07B	800	10.824 (5.344)					4.591 (2.309)	2.644 (1.502)	3.589 (2.491)	
	8	1321	11.775 (5.687)					4.861 (2.359)	2.864 (1.597)	4.050 (2.697)	
	09A	386	12.212 (6.105)					4.982 (2.375)	2.876 (1.621)	4.355 (2.988)	
	10A	321	12.879 (7.243)					5.202 (2.807)	3.016 (1.769)	4.660 (3.487)	
ST4	09B	369	11.756 (6.02)							4.772 (3.094)	6.984 (3.529)
	10B	315	11.514 (6.701)							4.663 (3.287)	6.851 (3.932)
	11	707	12.795 (6.181)							5.413 (3.059)	7.382 (3.775)

Table 1.7. Scaling Test Raw Scores, by Grade—Science

Scaling Test	Grade	N	Mean (SD) on ST	Mean (SD) of Grade Specific Items in Scaling Tests							
				G3	G4	G5	G6	G7	G8	G_EHS	G_ACT
ST1	3	1806	12.108 (5.049)	3.538 (1.515)	6.095 (2.909)	2.475 (1.699)					
	4	1551	14.453 (5.466)	3.983 (1.434)	7.362 (3.168)	3.108 (1.873)					
	05A	727	16.993 (5.945)	4.415 (1.286)	8.693 (3.484)	3.884 (2.129)					
ST2	05B	721	15.434 (6.267)			4.277 (2.080)	7.712 (3.247)	3.445 (1.852)			
	6	1514	15.687 (6.789)			4.327 (2.236)	7.814 (3.448)	3.546 (1.984)			
	07A	819	17.098 (6.836)			4.750 (2.245)	8.457 (3.440)	3.891 (1.984)			
ST3	07B	834	14.138 (6.306)					4.125 (1.910)	4.788 (2.266)	5.225 (3.302)	
	8	1426	15.243 (6.993)					4.391 (1.946)	5.100 (2.400)	5.752 (3.749)	
	09A	336	16.202 (8.015)					4.464 (2.141)	5.390 (2.728)	6.348 (4.195)	
	10A	298	16.319 (8.281)					4.470 (2.181)	5.366 (2.707)	6.483 (4.425)	
ST4	09B	328	14.137 (6.547)							6.747 (3.832)	7.390 (3.517)
	10B	320	15.272 (7.795)							7.481 (4.484)	7.791 (4.066)
	11	685	15.54 (7.576)							7.672 (4.232)	7.869 (4.138)

Figure 1.3. Scatter plots of average scale scores against scaling test raw scores, by grade

A. English

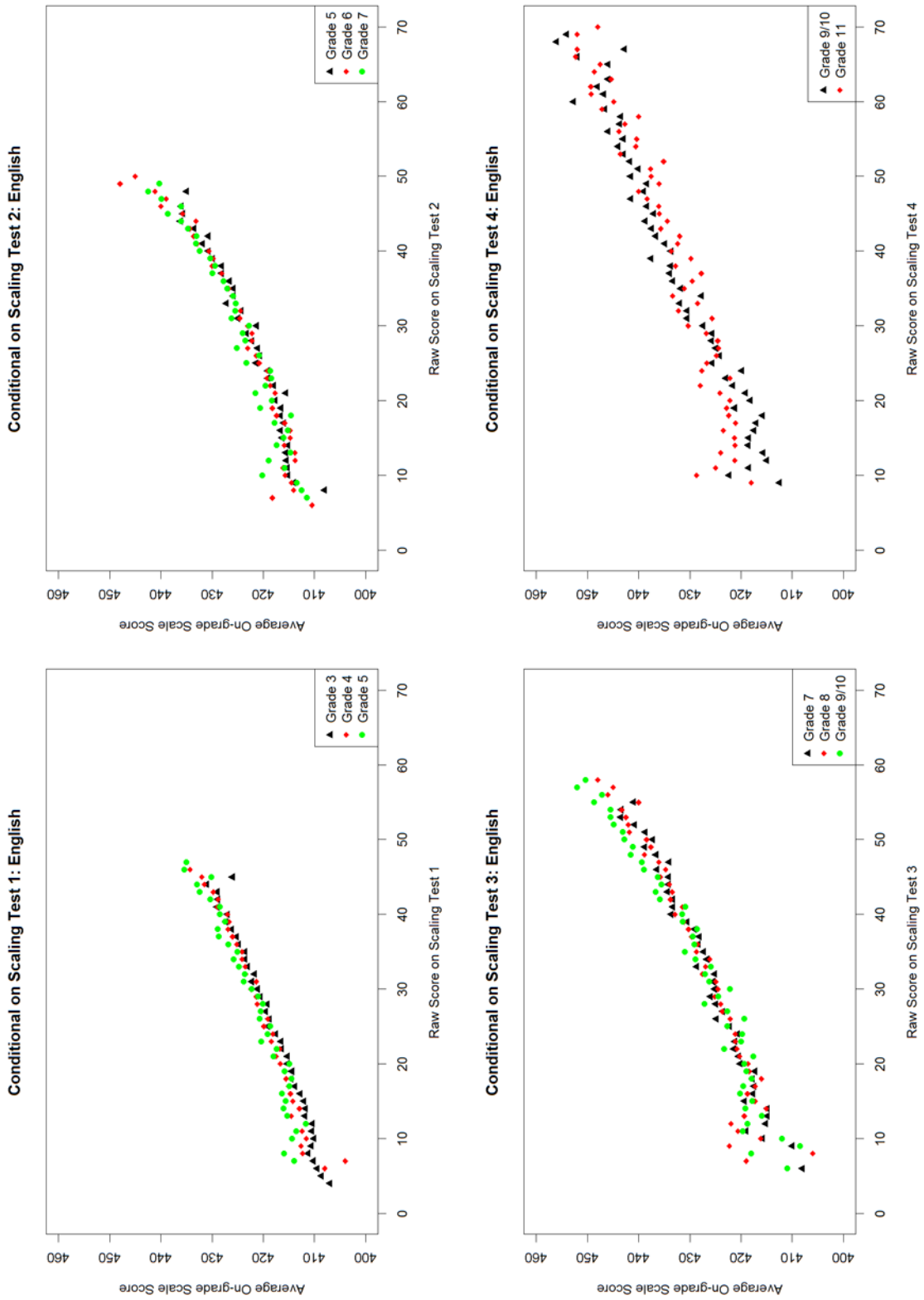


Figure 1.3. (continued)

B. Math

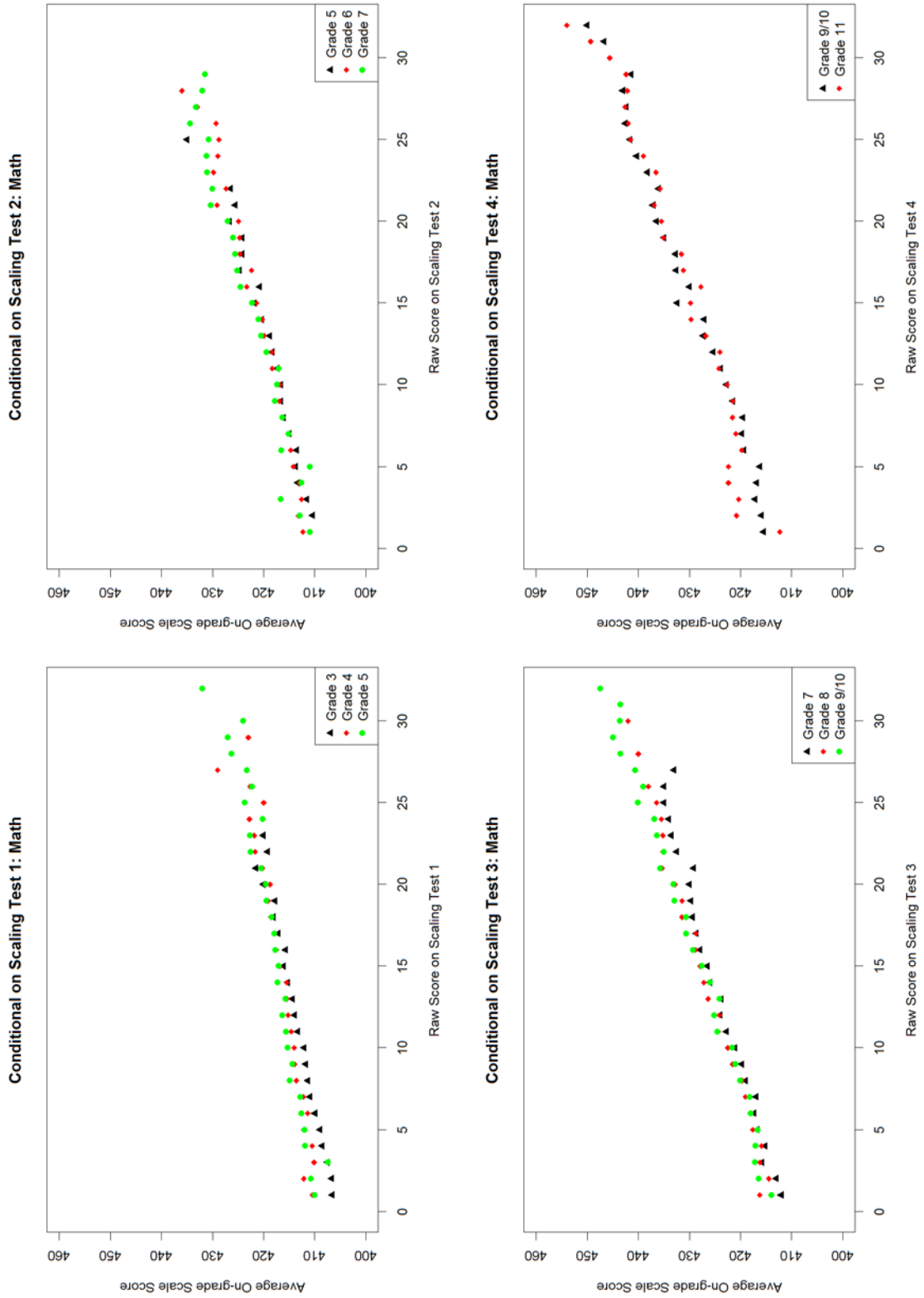


Figure 1.3. (continued)

C. Reading

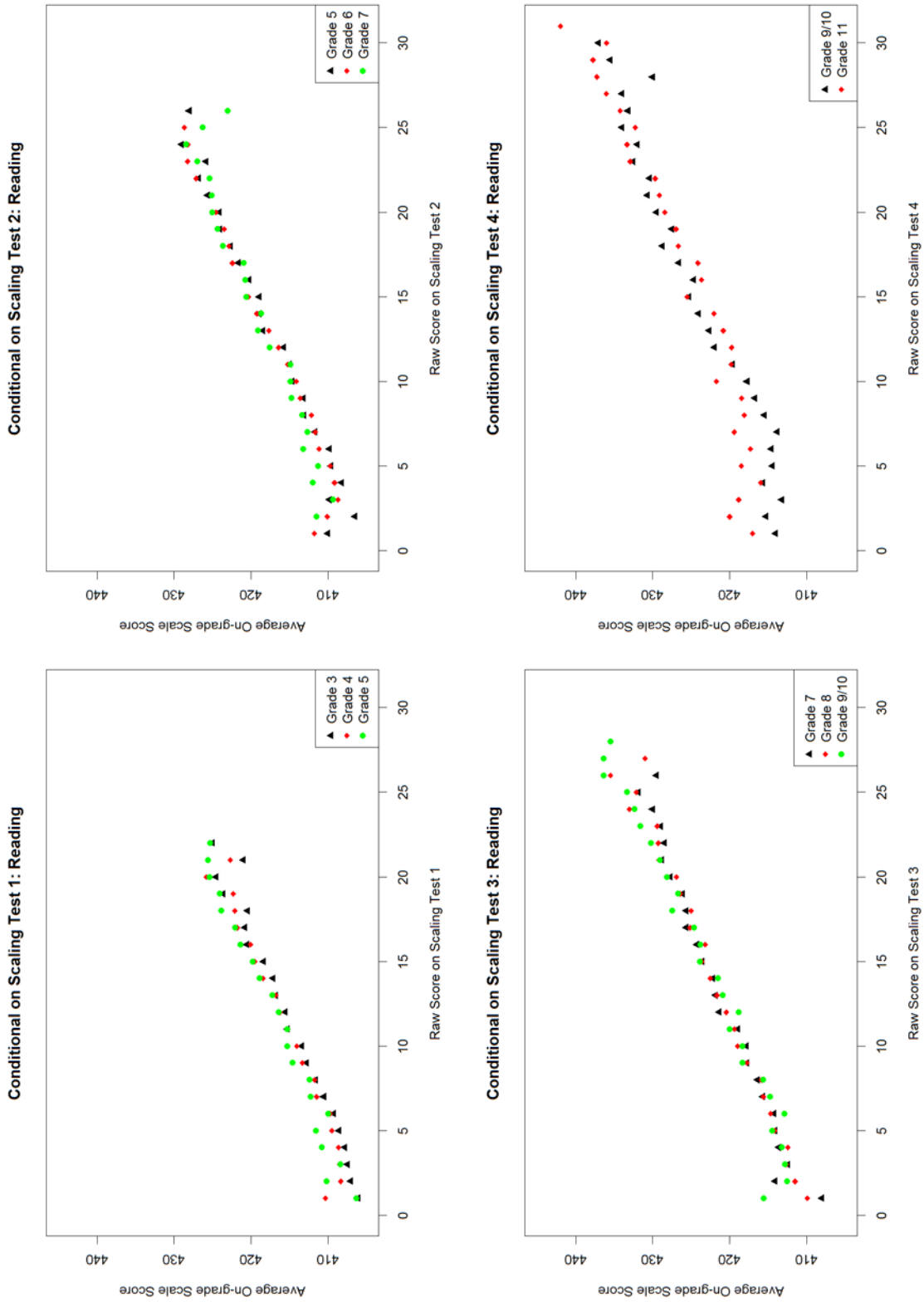
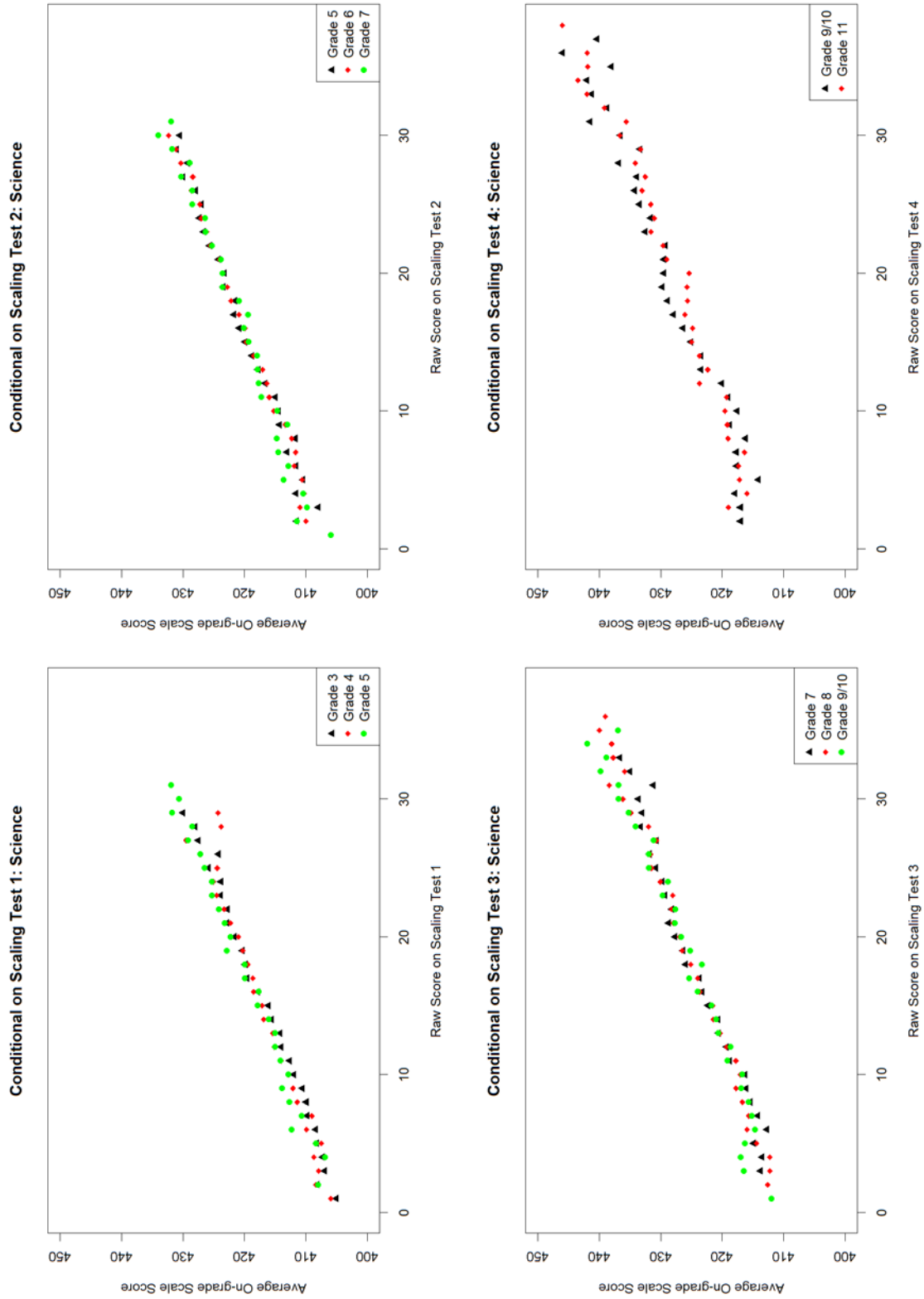


Figure 1.3. (continued)

D. Science

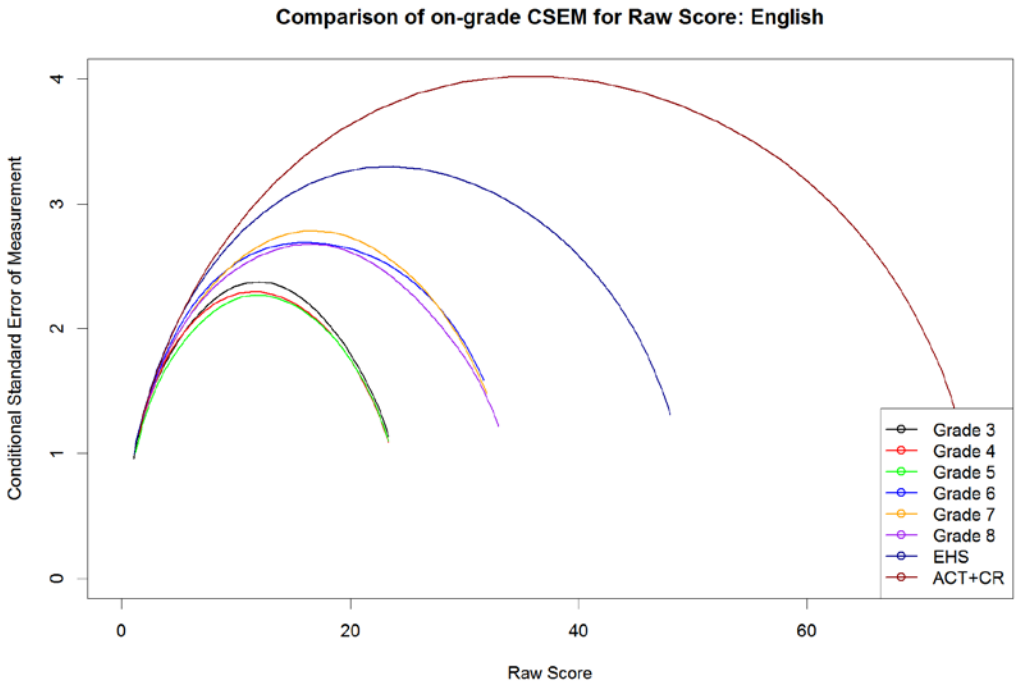


Evaluating the Constant CSEM Property

Figure 1.4 displays the CSEMs of raw scores and scale scores by subject. For most grades and subjects, the scale score CSEM curves are relatively flat along the scale range, especially when compared to the raw score CSEM, which show a typical inverted U-shape.

Figure 1.4. Conditional standard error of measurement, raw scores and scale scores

A. English



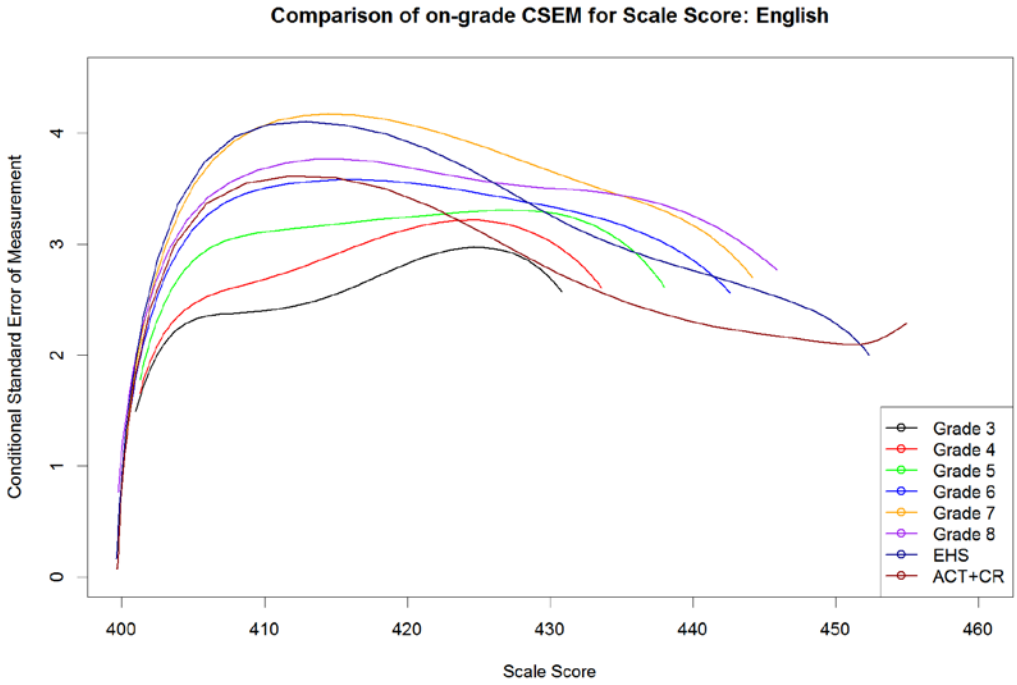
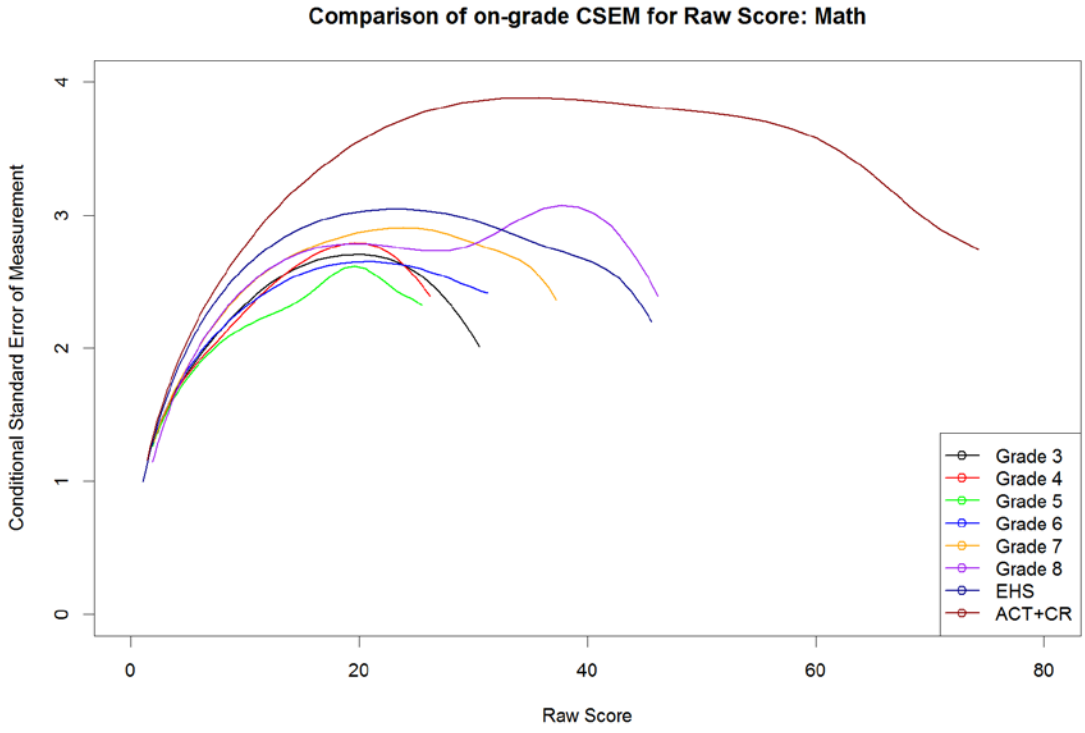


Figure 1.4. (continued)

B. Math



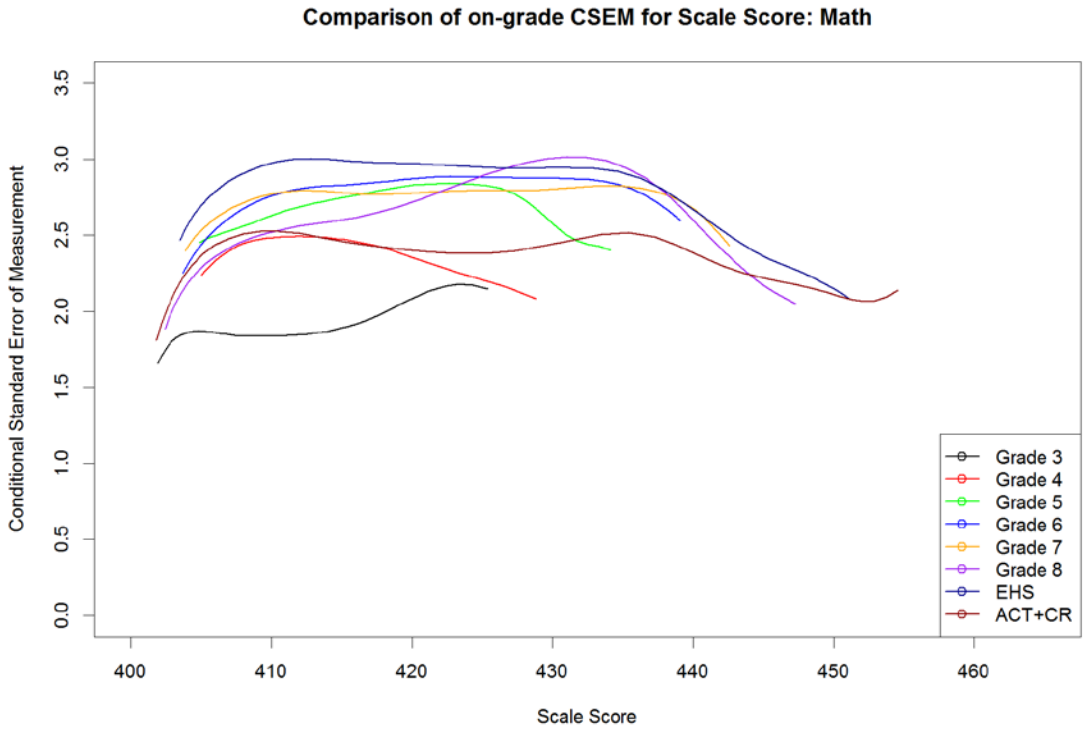
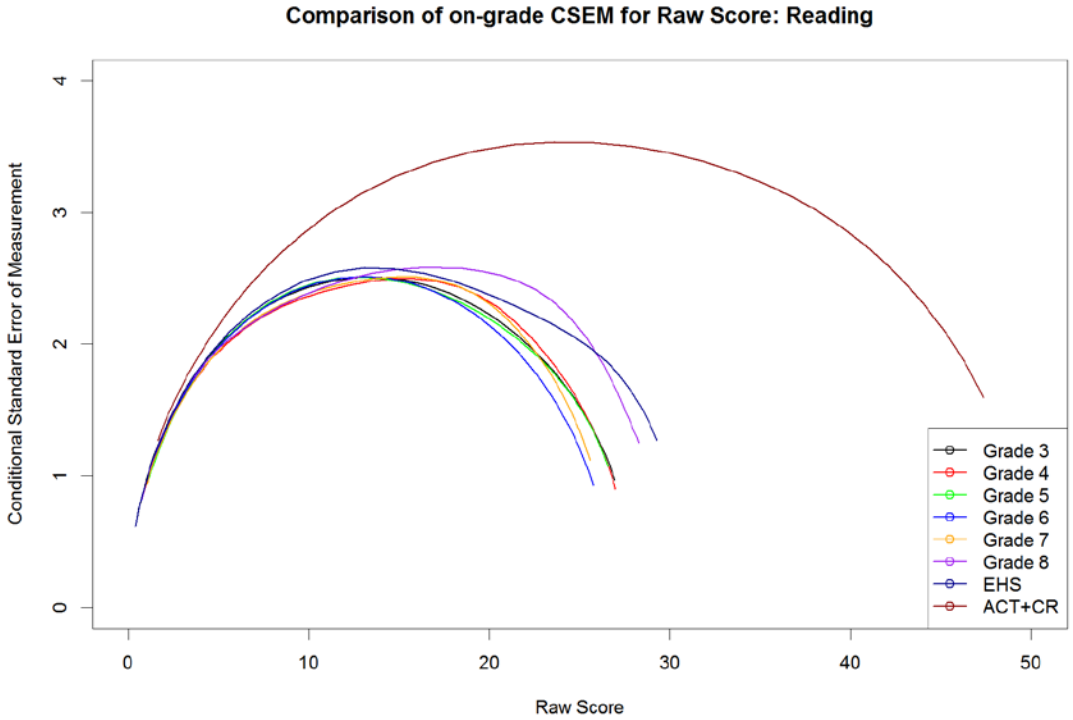


Figure 1.4. (continued)

C. Reading



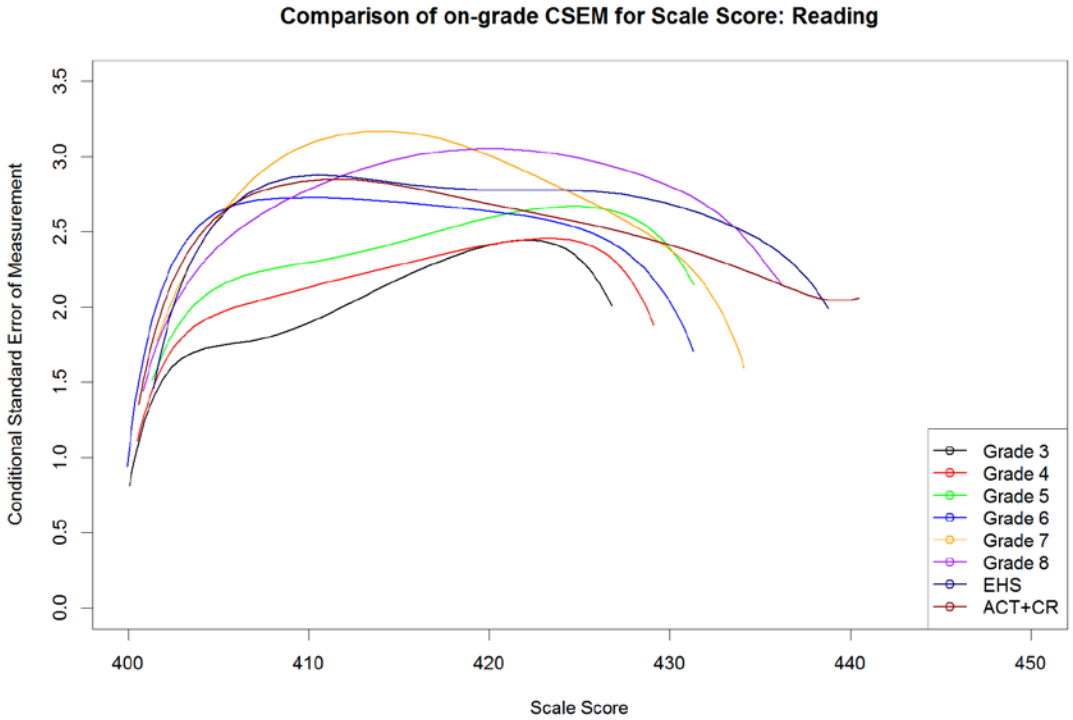
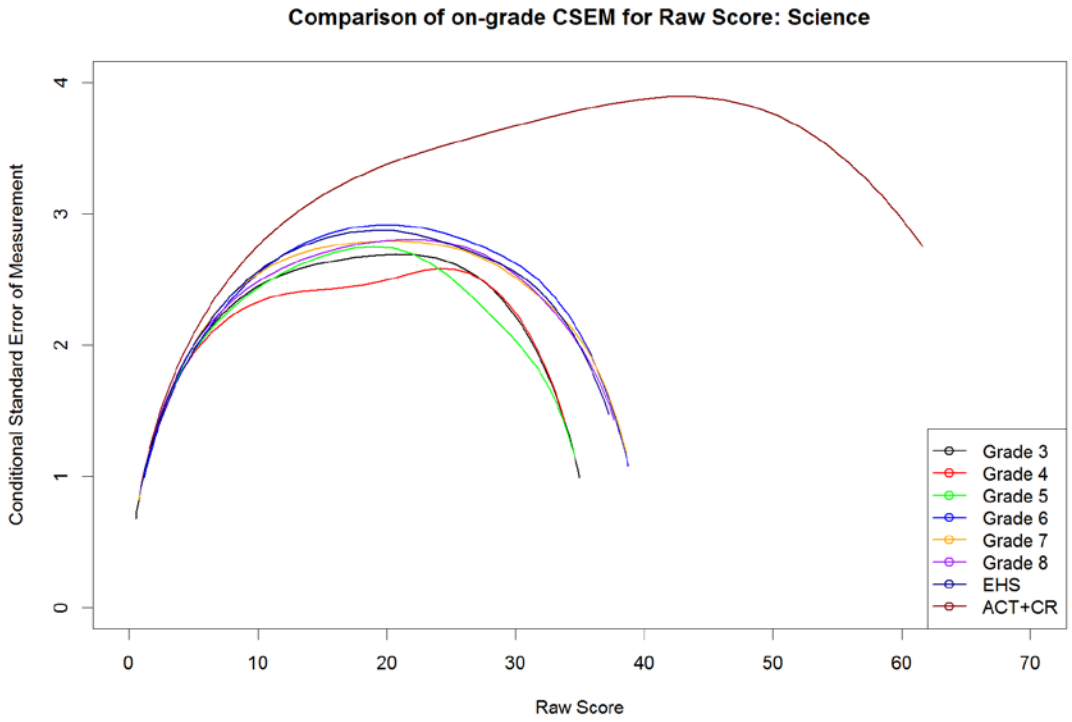
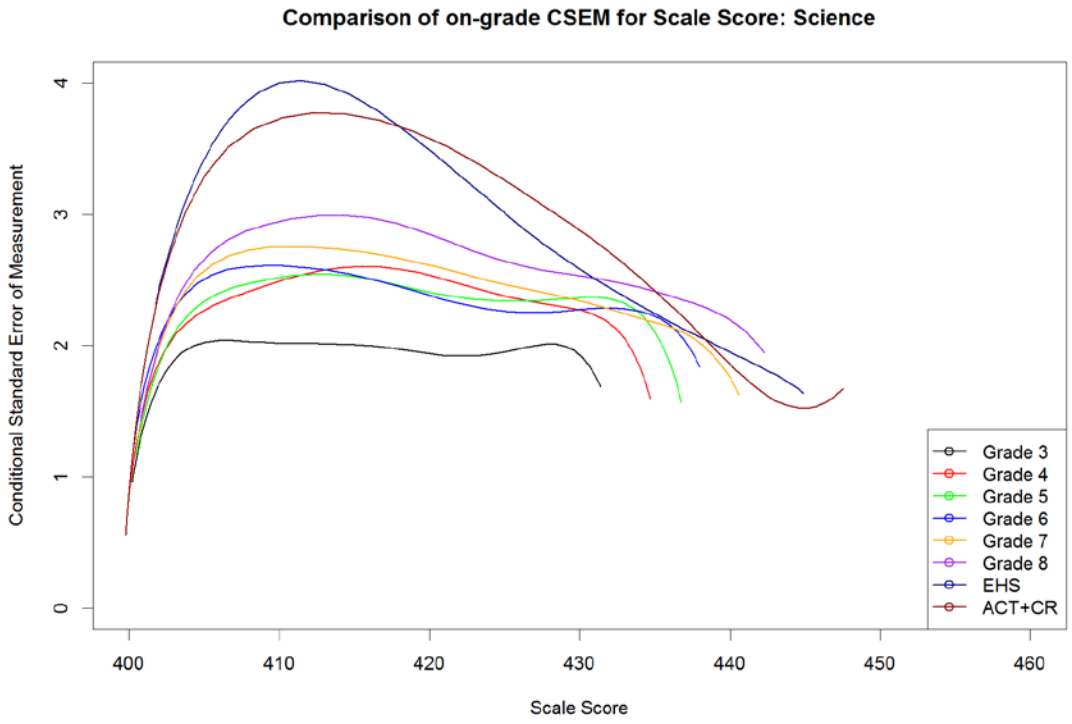


Figure 1.4. (continued)

D. Science





Growth Patterns

Effect sizes derived from comparing different grade levels across three types of scores were computed and compared. Figure 1.5 displays three sets of effect sizes, one derived from the final ACT Aspire scale scores on the on-grade test, one from the projected whole scaling test scores, and one from individual scaling test raw scores. Effect sizes derived from individual scaling test raw scores were computed between two groups taking the same scaling test, as described in the section describing method selection.

In general, the three sets of effect sizes for lower grades (3–8) are very similar. The patterns diverge in some cases for upper grades. One reason might be that grade 9 and 10 students were combined during the scaling analysis. Also, there were two sources of growth from grade 9 to grade 10: between 9A and 10A students who took ST3 and between 9B and 10B who took ST4. The effect size computed using the scaling test raw scores was the average of both sources. If the growth patterns are different from 9A to 10A than from 9B to 10B, this might lead to different effect sizes between 9 and 10 when combining the A and B groups.

Figure 1.5. Effect sizes computed from the final ACT Aspire scale, whole scaling test scores, and individual scaling test raw scores

English

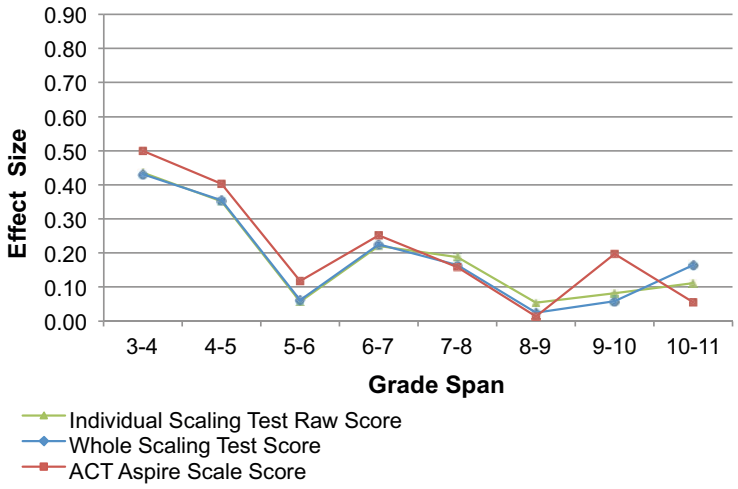
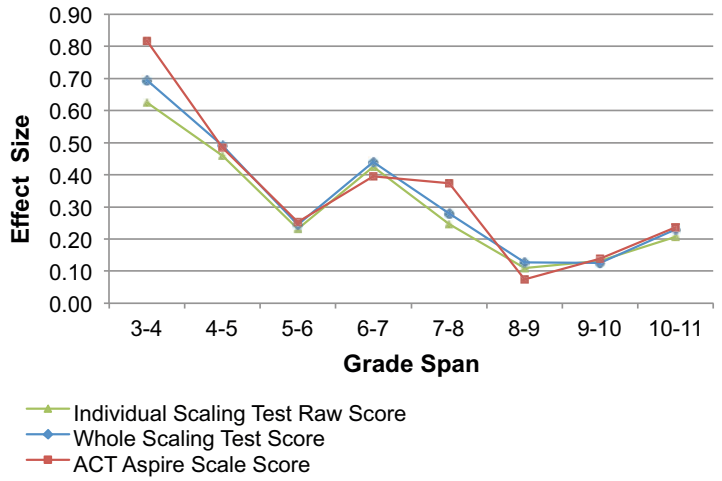
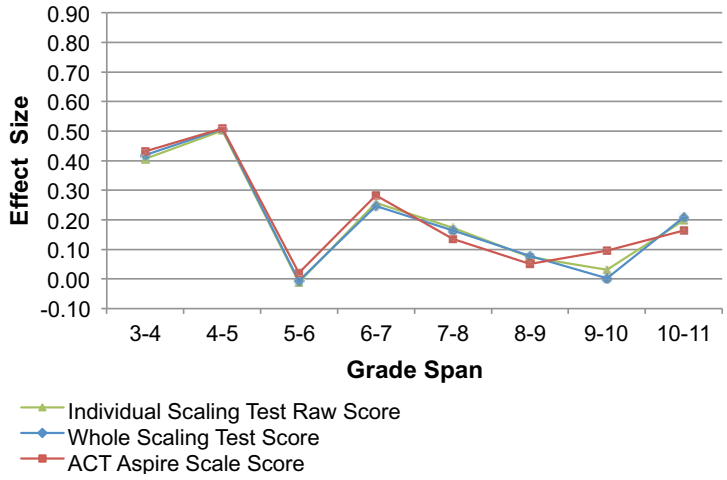


Figure 1.5. (continued)

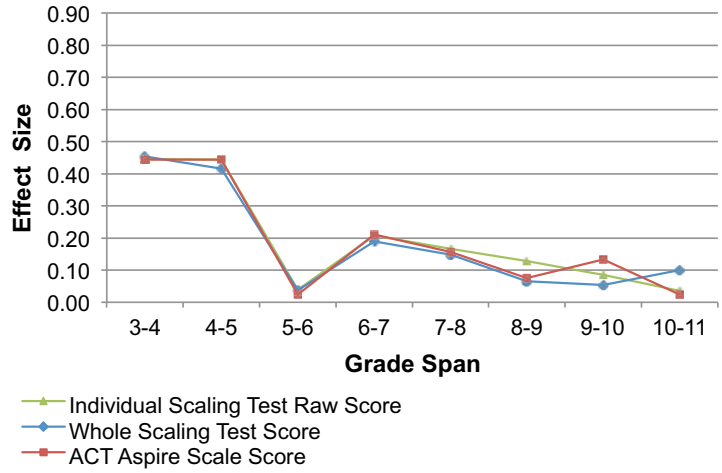
Math



Reading



Science



Tables 1.8–1.12 presents basic descriptive statistics of the final scale scores for all students in the scaling study. Note that the *n*-counts in these tables are larger than the groups used for the scaling analysis, where only students who took both scaling tests and on-grade tests were included. Since this is not longitudinal data and samples may not be representative in each grade, the decreasing means observed for certain subjects as grade increases may be explainable as a characteristic of the sample rather than a characteristic of the scale.[6] Future results from operational administrations with larger samples of highly motivated students taking ACT Aspire over multiple years will provide more robust estimates of longitudinal growth on the ACT Aspire scale.

Table 1.8. ACT Aspire English Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade

Grade	N	Mean	SD	Min	P10	P25	P50	P75	P90	P95	Max
3	4,277	416.50	6.14	403	409	412	416	421	424	428	435
4	3,825	419.65	6.23	402	411	415	420	423	428	430	438
5	4,301	421.98	7.04	403	413	417	421	427	431	435	442
6	4,683	422.56	8.09	400	412	416	423	428	433	437	448
7	4,762	424.84	8.56	400	413	419	425	430	436	439	450
8	3,609	426.38	8.94	401	415	421	426	433	438	442	452
9	2,415	425.48	10.76	400	413	417	424	433	440	445	456
10	2,258	429.07	11.59	400	414	419	429	439	445	447	456
11	2,240	429.15	11.64	400	414	420	429	438	445	448	460

Table 1.9. ACT Aspire Math Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade

Grade	N	Mean	SD	Min	P10	P25	P50	P75	P90	P95	Max
3	4,300	411.75	3.69	400	407	409	412	414	417	418	426
4	3,830	414.75	3.81	403	410	413	415	417	420	421	429
5	4,260	416.78	4.43	404	412	414	417	419	422	424	435
6	4,577	418.15	5.72	402	412	414	418	421	426	429	441
7	4,497	419.83	6.42	402	413	415	419	424	429	431	442
8	3,746	422.48	7.30	403	413	417	422	427	432	436	448
9	2,479	422.82	8.05	406	414	417	422	428	434	438	449
10	2,528	425.15	9.10	406	414	418	424	432	438	441	450
11	1,796	427.34	9.38	407	417	420	426	433	441	445	457

[6] Writing scale scores would not be expected to increase across grades due to how it was scaled. See "Scaling ACT Aspire Writing."

Table 1.10. ACT Aspire Reading Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade

Grade	N	Mean	SD	Min	P10	P25	P50	P75	P90	P95	Max
3	4,307	411.56	5.27	401	406	407	411	415	419	422	429
4	3,661	414.02	5.66	401	407	410	414	418	423	424	431
5	4,129	416.20	6.23	401	408	411	416	420	425	427	434
6	4,520	416.90	6.84	402	409	412	416	422	427	428	436
7	4,475	418.75	6.73	402	410	414	419	424	427	429	438
8	3,585	419.76	7.14	401	410	414	420	425	429	431	440
9	2,257	419.71	7.86	403	410	414	419	426	430	433	442
10	2,260	421.30	8.22	403	411	415	421	428	433	434	442
11	1,789	422.43	7.37	402	413	417	422	428	433	435	442

Table 1.11. ACT Aspire Science Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade

Grade	N	Mean	SD	Min	P10	P25	P50	P75	P90	P95	Max
3	4,214	413.89	5.99	401	407	409	414	418	422	424	433
4	3,571	416.29	6.53	400	407	412	416	421	425	427	435
5	3,903	419.05	6.78	401	410	414	420	424	427	430	438
6	4,642	418.76	7.54	400	409	412	419	424	429	431	440
7	4,756	420.49	7.65	401	410	414	421	426	431	432	441
8	3,544	422.14	7.91	401	412	416	422	427	433	435	446
9	2,167	422.65	8.22	402	412	417	421	429	434	437	447
10	2,314	424.43	8.89	402	414	417	424	431	436	439	449
11	1,717	424.59	9.04	400	412	417	425	431	437	439	449

Table 1.12. ACT Aspire Writing Scale Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade

Grade	N	Mean	SD	Min	P10	P25	P50	P75	P90	P95	Max
3	4,109	422.57	6.49	408	440	414	418	422	428	430	432
4	3,439	423.16	7.06	408	440	416	418	424	428	432	434
5	3,935	422.21	7.00	408	440	412	418	424	426	432	432
6	4,512	425.80	8.06	408	448	416	420	426	432	436	440
7	4,666	425.47	6.87	408	448	416	420	426	430	434	438
8	3,522	422.66	6.40	408	448	416	418	424	426	432	432
9	2,117	422.22	7.66	408	448	410	418	424	428	432	434
10	2,258	424.40	8.28	408	448	412	418	424	430	434	440
11	1,717	424.59	9.04	400	412	417	425	431	437	439	449

Table 1.13 presents the lowest obtainable scale scores (LOSS) and the highest obtainable scale scores (HOSS) of the ACT Aspire scales for all subjects and all grades. Note that English, Mathematics, Reading, and Science tests all have the same minimum scale score of 400 but the maximum scale scores vary across grade level within a subject. Writing tests all have the same minimum scale score of 408 but the maximum scale scores are 440 or 448, depending on grade level. Scaling of the ACT Aspire Writing test was conducted differently from the other subjects and is described in the next section.

Table 1.13. Lowest Obtainable Scale Scores (LOSS) and Highest Obtainable Scale Scores (HOSS)

Subject	Grade	LOSS	HOSS	Subject	Grade	LOSS	HOSS
English	3	400	435	**Science**	3	400	433
	4	400	438		4	400	436
	5	400	442		5	400	438
	6	400	448		6	400	440
	7	400	450		7	400	443
	8	400	452		8	400	446
	9	400	456		9	400	449
	10	400	456		10	400	449
Mathematics	3	400	434	**Writing**	3	408	440
	4	400	440		4	408	440
	5	400	446		5	408	440
	6	400	451		6	408	448
	7	400	453		7	408	448
	8	400	456		8	408	448
	9	400	460		9	408	448
	10	400	460		10	408	448
Reading	3	400	429				
	4	400	431				
	5	400	434				
	6	400	436				
	7	400	438				
	8	400	440				
	9	400	442				
	10	400	442				

Scaling ACT Aspire Writing

The ACT Aspire Writing scale scores are rubric-driven and based on four domains (Ideas and Analysis, Development and Support, Organization, and Language Use and Conventions), with each domain scored on a 1–5 or a 1–6 scale (5 in grades 3–5, 6 in grades 6–EHS). The rubrics become more complex as grade increases, however, and a domain score of 4 indicates expected performance at each grade in each rubric domain. The total raw writing score is the sum of the four raw domain scores and ranges from 4 to 20 for grades 3–5 and from 4 to 24 for grades 6–EHS. A common linear function shared by all grades was used to convert the total raw scores to three-digit ACT Aspire scale scores for the base writing forms. The lowest obtainable scale score for writing is 408 for all grades, and the highest obtainable scale score is 440 for grades 3–5 and 448 for grades 6–EHS. Table 1.13 lists the lowest and highest observable scale scores for writing by grade level.[7]

The scoring process for writing implies that Writing test scale scores do not have a vertical scale interpretation like those for scale scores obtained in the other subjects. Specifically, Writing scale scores across grade levels reflect performance on the rubric-driven expectations; the underlying performance needed to receive the same score is automatically adjusted to grade-level expectations. Consistency in scores across grade levels suggests the same performance relative to grade-level expectations. For example, if a student receives a score of 432 on the Writing test at grade 3 and a score of 432 at grade 8, it means that the student has achieved a level of performance that is consistently located relative to grade-level expectations. In other words, the same scores across grades 3 and 8 means that a student has grown to meet grade-level expectations. Table 1.12 lists the descriptive statistics for ACT Aspire Writing scale scores from the sample of students included in the scaling study.

[7] Additional descriptions of ACT Aspire Writing scale scores are provided in chapter 2.

CHAPTER 2

ACT Aspire Scores

ACT Aspire Subject Scores

ACT Aspire scale scores are reported from grade 3 through EHS in English, mathematics, reading, science and writing. Scale scores for each subject are provided on a three-digit scale. Scale score ranges for each grade and subject are provided in chapter 1. Summary descriptive statistics for the ACT Aspire subject scale scores from the 2013 scaling study are listed in tables 1.8–1.12.

ACT Aspire Composite Score

The ACT Aspire Composite score represents the overall performance on the English, Mathematics, Reading and Science tests.[8] It is calculated as the average of the scale scores on the four subjects rounded to an integer value (.5 rounds up). The Composite score is reported only to students taking grade 8 or EHS tests in all four subjects. Table 2.1 lists descriptive statistics for ACT Aspire Composite scores using data from the 2013 scaling study.

Table 2.1. ACT Aspire Composite Score Descriptive Statistics Based on the Entire Sample of the Scaling Study, by Grade

Grade	N	Mean	SD	Min	P10	P25	P50	P75	P90	P95	Max
8	2945	423.02	6.84	406	442	414	418	423	428	432	435
9	1641	423.99	7.93	408	447	414	418	423	430	435	438
10	1649	426.28	8.53	409	447	415	419	426	433	438	440

[8] Note that the writing test is not included in the composite score.

Averaging four ACT Aspire subject test scale scores to obtain a Composite score implies that each test is contributing equally to the Composite score. The weights used to calculate the Composite (in this case, .25) are often referred to as nominal weights. Other definitions of the contribution of a test score to a composite score may be more useful. For example, Wang and Stanley (1970) described effective weights as an index of the contribution of a test score to a composite score. Specifically, the contribution of a test score is defined as the sum of the covariances between the test score and all components contributing to the Composite score. These contributions can be summed over tests and then each can be divided by its sum to arrive at proportional effective weights. Proportional effective weights are referred to as effective weights here.

With nominal weights of .25 for each test, the effective weights can be used to verify that the nominal weight interpretation of Composite scores (i.e., composite as an equal weighted combination of contributing scores) is reasonable. Wang and Stanley (1970) state that variables will rarely have equal effective weights, unless explicitly designed to do so. Therefore, the effective weights would need to deviate substantially and consistently from nominal weights to justify applying different weights or a different interpretation of weights.

The effective weights using spring 2013 scale scores are shown in table 2.2 for the ACT Aspire Composite scores and range from .14 to .32. The Mathematics tests at grades 3–6 appeared to have smaller effective weights compared to English, Reading, and Science due to relatively smaller variances and covariances. However, the equal nominal weights appeared justifiable for ACT Aspire Composite scores.

Table 2.2. ACT Aspire Composite Score Effective Weights Assuming Equal Nominal Weights

Grade	English	Mathematics	Reading	Science
3	.29	.16	.25	.29
4	.28	.14	.27	.31
5	.29	.15	.26	.29
6	.29	.18	.25	.28
7	.29	.21	.23	.27
8	.29	.23	.23	.26
9	.32	.23	.22	.24
10	.31	.24	.22	.24

Table 2.3. ACT Aspire ELA Effective Weights Assuming Equal Nominal Weights

Grade	English	Reading	Writing
3	.35	.30	.35
4	.33	.31	.36
5	.36	.32	.33
6	.36	.30	.33
7	.41	.31	.29
8	.42	.32	.26
9	.43	.30	.27
10	.43	.30	.27

ACT Aspire ELA Score

The ACT Aspire ELA score represents the overall performance on the English, Reading and Writing tests. It is calculated as the average of the scale scores on the three subjects rounded to an integer value (.5 rounds up). ELA scores are provided to students at all grade levels between 3 and EHS but are provided only when a student obtains scale scores for all three subject tests at the same grade level. Nominal weights for ELA are .33 (equal weights for each of the three contributing subject scores). The effective weights of spring 2013 scale scores are listed in table 2.3 and ranged from .26 to .43. English appeared to contribute more to the effective weights for grades 7–10, with weights greater than .40, compared to Writing, where weights dipped below .30. However, the effective weights did not deviate far from the equal nominal weights for ELA scores.

ACT Aspire STEM Score

The ACT Aspire STEM score represents the overall performance on the Mathematics and Science tests. It is calculated as the average of the scale scores on the two subjects rounded to an integer value (.5 rounds up). The STEM score is provided for students at all grade levels between 3 and EHS but only when a student obtains scale scores for Mathematics and Science tests at the same grade level. Nominal weights for STEM are .5 (equal weights for each of the two contributing subject scores). The effective weights of spring 2013 scale scores are listed in table 2.4 and ranged from .33 to .67. Similar to the Composite scores, the Mathematics scores had smaller variances and covariances compared to Science, particularly for grades 3, 4, and 5, which contributed to smaller effective weights. Despite the observed differences at lower grades, where Science contributed nearly double to the STEM score variance compared to Mathematics, the effective weights for grades 6–10 were progressively more similar to the nominal weights.

Table 2.4. ACT Aspire STEM Effective Weights Assuming Equal Nominal Weights

Grade	Mathematics	Science
3	.36	.64
4	.33	.67
5	.37	.63
6	.42	.58
7	.45	.55
8	.48	.52
9	.50	.50
10	.50	.50

Reporting Categories

Student performance is also described in terms of the ACT Aspire reporting categories. Score reports describe the percent and number of points students earn out of the total number of points possible in each reporting category. Descriptions of reporting categories by subject are included in *Technical Bulletin #1* (ACT 2014b).

Progress with Text Complexity

The ACT Aspire Reading test includes a Progress with Text Complexity indicator ("Yes" or "No") to indicate whether students have made sufficient progress in reading increasingly complex texts (see ACT 2014b). This indicator is based on a set of items from the ACT Aspire Reading test that are judged as indicative of performance when reading and understanding increasingly complex texts. It is calculated by comparing the percent correct on the items to an empirically defined cut score. Regression was used to predict the percent correct on the set of items indicative of progress with text complexity using the reading scale score at the reading benchmark (see chapter 5 for a description of the reading benchmark).

CHAPTER 3

ACT Aspire Norms

ACT Aspire norms are defined as the cumulative percent of students scoring at or below a given score in the norm sample. The norm sample is a reference sample of students taking the ACT Aspire tests. Norms tables list the cumulative percent of students who are at or below each scale score point in the norm sample. For example, a cumulative percent of 50 for a scale score of 420 on the Mathematics test means that 50% of students in the norm group achieved a scale score of 420 or below on that test.

In 2014, normative information was provided based on student samples from 2013 special study data and 2014 operational data. Starting in 2015, ACT Aspire will begin reporting three-year rolling norms, similar to The ACT. Rolling norms will include student samples from the three most recent years of ACT Aspire test administrations.

ACT Aspire norms are national, with a broad representation across the country, but they are not nationally representative norms, since they have not been statistically weighted to more closely mimic the national distribution of demographics. ACT Aspire includes national norms for grades 3 through EHS in English, mathematics, reading, science and writing. Norm group demographic characteristics, including gender, state, and race/ethnicity, for each grade and subject area are included in tables 3.1–3.5. The 2014 ACT Aspire norms are listed in tables 3.6–3.10.

Table 3.1. 2014 ACT Aspire English Norm Group Demographics

	Grade (%)							
	3 (n = 36,139)	**4** (n = 36,353)	**5** (n = 36,890)	**6** (n = 35,258)	**7** (n = 35,618)	**8** (n = 37,796)	**9** (n = 23,096)	**10** (n = 19,417)
Gender								
F	45.24	45.29	46.12	44.99	44.49	45.41	44.34	44.26
M	47.21	47.22	46.89	46.11	45.61	45.72	43.31	43.59
No response	7.56	7.48	6.99	8.90	9.90	8.87	12.36	12.15
State								
AL	66.38	67.74	65.16	61.33	56.29	51.44	7.13	6.44
AR	0.05	0.05	0.04	0.07	0.04	0.06	0.07	0.09
AZ	—	—	—	—	—	0.15	0.31	0.26
CA	0.28	0.24	2.94	2.99	0.33	0.24	1.26	1.59
CO	2.42	2.28	2.03	1.70	1.62	1.91	2.22	1.98
CT	0.16	0.16	0.20	0.18	0.12	0.19	1.27	1.34
FL	0.06	0.11	—	—	—	0.10	1.02	0.62
GA	0.08	0.09	0.12	0.10	—	—	—	—
IA	0.54	0.58	0.54	0.35	0.52	0.33	0.23	1.12
IL	1.54	1.47	1.60	1.79	1.87	1.66	3.97	4.29
IN	1.01	1.10	1.02	2.22	1.84	0.64	1.13	1.73
KS	1.66	1.64	1.51	2.84	2.54	1.64	2.55	2.44
KY	4.91	4.98	4.03	3.40	3.82	2.66	5.94	2.42
LA	2.17	2.07	1.99	2.05	1.66	1.55	3.16	2.98
MA	—	—	—	—	—	—	—	0.09
MI	3.65	2.96	3.59	4.75	8.41	17.28	22.38	26.19
MN	0.32	0.31	0.34	0.48	0.38	0.37	—	0.52
MO	1.84	1.44	1.93	1.65	1.51	1.37	2.93	4.11
MS	2.27	2.64	2.59	2.77	2.65	2.07	2.55	1.48
ND	—	—	—	—	—	0.23	0.16	0.31
NE	—	—	—	—	—	—	1.59	1.89
NJ	—	—	—	—	—	—	0.44	—
NM	—	—	—	—	—	—	0.57	—
NV	—	—	—	—	—	0.09	0.06	0.07
OH	1.48	1.24	1.14	1.61	3.76	1.74	9.30	5.28

Note: — indicates no students tested.

Table 3.1. (continued)

	Grade (%)							
	3 (*n* = 36,139)	**4** (*n* = 36,353)	**5** (*n* = 36,890)	**6** (*n* = 35,258)	**7** (*n* = 35,618)	**8** (*n* = 37,796)	**9** (*n* = 23,096)	**10** (*n* = 19,417)
OK	0.70	0.91	0.69	1.14	1.08	1.03	0.14	0.09
PA	—	—	—	—	—	—	1.13	1.18
SC	3.61	3.57	3.63	3.73	3.66	3.63	1.69	1.99
SD	—	—	0.20	—	—	0.21	—	0.25
TN	1.31	1.43	1.42	1.45	1.84	1.95	4.30	4.99
TX	1.96	1.16	1.60	1.32	1.90	2.07	2.32	5.04
UT	0.79	0.82	0.80	0.96	1.34	0.96	0.71	—
VA	—	—	—	0.08	—	0.25	—	—
WI	0.82	1.01	0.89	1.01	2.74	4.11	19.16	18.98
No response	—	—	—	0.01	0.06	0.08	0.31	0.27
Race/ Ethnicity								
Black/African American	22.62	22.41	22.43	21.47	20.97	18.37	9.07	8.49
American Indian/Alaska Native	1.07	1.03	0.92	1.19	0.82	0.88	0.52	0.64
White	44.81	45.24	43.64	41.79	42.25	44.53	35.21	38.01
Hispanic/ Latino	5.47	5.27	5.29	6.67	6.52	5.62	5.19	4.65
Asian	1.68	1.62	1.59	1.28	1.42	1.44	1.40	2.02
Native Hawaiian/ Other Pacific Islander	0.58	0.52	0.91	0.78	0.55	0.40	0.48	0.37
Two or more races	0.41	0.32	0.32	0.41	0.61	0.45	0.51	0.78
No response	23.35	23.59	24.89	26.42	26.85	28.31	47.63	45.04

Note: — indicates no students tested.

Table 3.2. 2014 ACT Aspire Mathematics Norm Group Demographics

	Grade (%)							
	3 (n = 76,861)	**4** (n = 76,012)	**5** (n = 77,083)	**6** (n = 75,651)	**7** (n = 77,891)	**8** (n = 80,763)	**9** (n = 23,011)	**10** (n = 19,441)
Gender								
F	47.15	47.00	47.30	46.91	46.46	47.50	44.60	44.04
M	48.92	49.08	48.96	48.50	48.62	48.19	43.26	43.56
No response	3.93	3.92	3.74	4.59	4.92	4.32	12.14	12.40
State								
AL	83.98	84.62	82.65	81.73	80.05	76.23	5.53	4.99
AR	0.02	0.02	0.02	0.03	0.02	0.03	0.07	0.09
AZ	—	—	—	—	—	0.07	0.30	0.26
CA	0.13	0.11	1.33	1.37	0.11	0.09	1.24	1.62
CO	1.15	1.10	0.98	0.81	0.74	0.89	2.23	2.06
CT	0.06	0.09	0.09	0.02	0.06	0.08	1.27	1.33
FL	0.03	0.05	—	—	—	0.04	0.98	0.59
GA	0.04	0.04	0.06	0.05	—	—	—	—
IA	0.26	0.28	0.26	0.16	0.24	0.15	0.24	1.15
ID	—	—	—	—	—	0.09	—	—
IL	0.74	0.70	0.76	0.85	0.86	0.79	4.05	3.93
IN	0.52	0.52	0.49	1.06	0.82	0.37	1.02	1.91
KS	0.81	0.77	0.73	1.20	1.00	0.75	2.88	2.74
KY	2.25	2.38	1.93	1.52	1.72	1.28	5.59	2.41
LA	1.03	0.99	0.95	0.95	0.76	0.73	3.09	2.95
MA	—	—	—	—	—	—	—	0.09
MI	1.77	1.44	1.88	2.79	4.10	8.67	22.38	26.31
MN	0.15	0.15	0.16	0.29	0.01	0.17	—	0.53
MO	0.85	0.69	0.91	0.72	0.72	0.64	2.89	3.37
MS	1.11	1.26	1.24	1.32	1.09	0.98	1.79	1.50
ND	—	—	—	—	—	0.11	0.17	0.31
NE	—	—	—	—	—	—	1.60	1.89
NJ	—	—	—	—	—	—	0.44	—
NM	—	—	—	—	—	—	0.59	—
NV	0.01	0.01	0.01	—	—	0.04	0.06	0.07

Note: — indicates no students tested.

Table 3.2. (continued)

	Grade (%)							
	3 (*n* = 76,861)	**4** (*n* = 76,012)	**5** (*n* = 77,083)	**6** (*n* = 75,651)	**7** (*n* = 77,891)	**8** (*n* = 80,763)	**9** (*n* = 23,011)	**10** (*n* = 19,441)
OH	0.69	0.56	0.54	0.75	1.65	0.94	9.51	5.92
OK	0.43	0.43	0.33	0.54	0.54	0.50	0.17	0.09
PA	—	—	—	—	—	—	1.00	1.18
SC	1.69	1.71	1.73	1.74	1.66	1.68	2.05	1.99
SD	—	—	0.09	—	—	0.10	—	0.25
TN	0.62	0.67	0.68	0.68	0.84	0.88	4.24	4.98
TX	0.90	0.57	1.39	0.56	1.07	1.04	2.63	5.12
UT	0.37	0.38	0.36	0.35	0.66	0.46	0.72	—
VA	—	—	—	0.04	—	0.12	—	—
WI	0.40	0.45	0.42	0.47	1.25	2.04	20.98	20.11
No response	—	—	—	0.00	0.03	0.04	0.30	0.27
Race/ Ethnicity								
Black/African American	26.63	27.08	27.02	27.63	27.71	26.77	8.14	8.29
American Indian/Alaska Native	0.84	0.84	0.83	0.93	0.78	0.80	0.65	0.67
White	50.40	50.64	50.00	49.05	49.75	50.61	35.77	36.98
Hispanic/ Latino	6.17	5.96	5.70	5.81	5.64	5.12	5.34	4.92
Asian	1.48	1.50	1.48	1.25	1.31	1.35	1.40	2.01
Native Hawaiian/ Other Pacific Islander	0.31	0.28	0.47	0.36	0.28	0.22	0.56	0.38
Two or more races	0.21	0.16	0.16	0.18	0.32	0.22	0.53	0.78
No response	13.96	13.54	14.34	14.79	14.21	14.91	47.63	45.97

Note: — indicates no students tested.

Table 3.3. 2014 ACT Aspire Reading Norm Group Demographics

	Grade (%)							
	3 (n = 76,817)	**4** (n = 75,577)	**5** (n = 76,316)	**6** (n = 74,862)	**7** (n = 77,497)	**8** (n = 80,077)	**9** (n = 22,711)	**10** (n = 19,060)
Gender								
F	47.17	47.10	47.34	46.93	46.59	47.48	44.57	43.92
M	49.05	49.12	49.06	48.56	48.84	48.20	43.15	43.48
No response	3.78	3.78	3.60	4.51	4.56	4.33	12.28	12.60
State								
AL	84.13	85.03	83.56	82.45	80.36	76.93	5.70	5.70
AR	0.02	0.02	0.02	0.03	0.02	0.03	0.07	0.09
AZ	—	—	—	—	—	0.07	0.32	0.27
CA	0.12	0.10	1.43	1.39	0.19	0.05	1.23	1.58
CO	1.15	1.10	0.98	0.81	0.75	0.91	2.23	2.03
CT	0.04	0.06	0.09	0.09	0.02	0.10	1.16	1.36
FL	0.03	0.05	—	—	—	0.05	1.06	0.63
GA	0.04	0.04	0.06	0.05	—	—	—	—
IA	0.26	0.28	0.26	0.17	0.24	0.15	0.24	1.17
ID	—	—	—	—	—	0.09	—	—
IL	0.73	0.70	0.77	0.85	0.83	0.79	4.12	4.47
IN	0.49	0.41	0.42	1.02	0.82	0.30	0.78	1.21
KS	0.77	0.74	0.67	1.18	1.00	0.77	2.29	2.46
KY	2.24	2.39	1.95	1.50	1.71	1.03	5.14	1.41
LA	1.03	1.00	0.96	0.94	0.76	0.73	3.05	2.90
MA	—	—	—	—	—	—	—	0.09
MI	1.83	1.48	1.85	2.26	3.88	8.46	22.80	26.78
MN	0.15	0.15	0.17	0.26	0.04	0.18	—	0.52
MO	0.85	0.70	0.89	0.75	0.71	0.55	2.90	3.81
MS	1.23	1.27	1.23	1.34	1.09	0.94	2.49	1.45
ND	—	—	—	—	—	0.11	0.17	0.31
NE	—	—	—	—	—	—	1.62	1.92
NJ	—	—	—	—	—	—	0.43	—
NM	—	—	—	—	—	—	0.61	—

Note: — indicates no students tested.

Table 3.3. (continued)

	Grade (%)							
	3 (*n* = 76,817)	**4** (*n* = 75,577)	**5** (*n* = 76,316)	**6** (*n* = 74,862)	**7** (*n* = 77,497)	**8** (*n* = 80,077)	**9** (*n* = 22,711)	**10** (*n* = 19,060)
NV	0.01	0.01	0.01	—	—	0.04	0.06	0.07
OH	0.69	0.59	0.55	0.65	1.75	0.93	9.70	5.67
OK	0.42	0.38	0.27	0.53	0.54	0.48	0.13	0.09
PA	—	—	—	—	—	—	0.96	1.23
SC	1.69	1.71	1.76	1.76	1.67	1.72	2.09	2.00
SD	—	—	0.09	—	—	0.10	—	0.27
TN	0.63	0.68	0.69	0.69	0.85	0.91	4.38	4.99
TX	0.92	0.56	0.75	0.62	0.85	0.94	2.27	5.24
UT	0.17	0.14	0.13	0.22	0.62	0.45	0.70	—
VA	—	—	—	0.03	—	0.12	—	—
WI	0.39	0.41	0.43	0.40	1.26	2.04	20.99	19.99
No response	—	—	—	0.00	0.03	0.04	0.32	0.27
Race/ Ethnicity								
Black/African American	26.87	27.18	27.22	27.82	27.90	26.92	8.44	7.96
American Indian/Alaska Native	0.82	0.83	0.82	0.92	0.77	0.80	0.66	0.71
White	50.30	50.77	50.01	49.17	49.74	50.54	35.08	36.85
Hispanic/ Latino	6.14	5.86	5.56	5.79	5.63	5.00	5.03	4.71
Asian	1.45	1.48	1.42	1.24	1.34	1.30	1.36	2.00
Native Hawaiian/ Other Pacific Islander	0.26	0.23	0.41	0.31	0.22	0.22	0.53	0.38
Two or more races	0.21	0.15	0.15	0.18	0.33	0.18	0.55	0.80
No response	13.94	13.50	14.41	14.57	14.09	15.04	48.35	46.59

Note: — indicates no students tested.

Table 3.4. 2014 ACT Aspire Science Norm Group Demographics

	Grade (%)							
	3 (n = 33,076)	**4** (n = 34,135)	**5** (n = 33,317)	**6** (n = 31,928)	**7** (n = 32,618)	**8** (n = 35,162)	**9** (n = 22,238)	**10** (n = 18,616)
Gender								
F	45.24	44.98	45.81	44.92	44.15	45.35	44.29	44.08
M	46.72	47.18	46.59	45.72	45.37	45.31	43.05	43.60
No response	8.04	7.84	7.59	9.36	10.48	9.34	12.66	12.32
State								
AL	63.30	66.95	62.04	57.38	52.50	49.04	5.81	5.23
AR	0.05	0.05	0.05	0.07	0.05	0.06	0.07	0.09
AZ	—	—	—	—	—	0.16	0.31	0.27
CA	0.24	0.24	3.16	3.19	0.29	0.18	0.36	1.21
CO	2.68	2.31	2.13	1.78	1.64	2.00	2.31	2.19
CT	0.15	0.19	0.20	0.19	0.13	0.17	1.17	1.39
FL	0.07	0.12	—	—	—	0.11	0.99	0.61
GA	0.08	0.09	0.14	0.12	—	—	—	—
IA	0.60	0.62	0.61	0.39	0.46	0.26	0.24	1.20
IL	1.76	1.58	1.72	1.90	2.06	1.81	3.98	4.26
IN	0.93	0.91	0.80	2.27	1.91	0.71	0.88	1.62
KS	1.86	1.71	1.67	3.16	2.80	1.76	2.32	2.35
KY	5.31	5.59	4.78	3.45	4.06	2.92	5.27	1.44
LA	2.36	2.20	2.19	2.22	1.74	1.67	3.11	3.01
MA	—	—	—	—	—	—	—	0.09
MI	3.99	2.74	4.35	6.12	9.31	17.46	22.21	27.14
MN	0.34	0.32	0.37	1.02	0.42	0.39	—	0.56
MO	1.85	1.50	1.91	1.84	1.67	1.46	2.76	3.24
MS	2.84	2.55	2.58	3.11	2.45	2.10	2.62	1.63
ND	—	—	—	—	—	0.24	0.17	0.32
NE	—	—	—	—	—	—	1.65	1.97
NJ	—	—	—	—	—	—	0.44	—
NM	—	—	—	—	—	—	0.60	—
NV	—	—	—	—	—	0.10	0.06	0.07

Note: — indicates no students tested.

Table 3.4. (continued)

	Grade (%)							
	3 (*n* = 33,076)	**4** (*n* = 34,135)	**5** (*n* = 33,317)	**6** (*n* = 31,928)	**7** (*n* = 32,618)	**8** (*n* = 35,162)	**9** (*n* = 22,238)	**10** (*n* = 18,616)
OH	1.68	1.27	1.32	1.51	4.43	2.07	9.92	6.53
OK	0.92	0.94	0.77	1.25	1.28	1.10	0.13	0.10
PA	—	—	—	—	—	—	0.98	1.26
SC	3.91	3.79	4.00	4.12	3.98	3.89	1.81	2.05
SD	—	—	0.22	—	—	0.23	—	0.25
TN	1.42	1.51	1.56	1.61	2.18	2.06	4.46	5.12
TX	2.12	1.27	1.78	1.37	2.07	2.44	2.77	4.79
UT	0.68	0.63	0.65	0.86	1.54	1.03	0.71	—
VA	—	—	—	0.09	—	0.27	—	—
WI	0.89	0.90	1.00	0.98	2.96	4.22	21.55	19.72
No response	—	—	—	0.01	0.07	0.09	0.32	0.28
Race/ Ethnicity								
Black/African American	21.71	22.31	23.33	21.86	22.99	18.78	8.16	7.71
American Indian/Alaska Native	0.98	0.86	0.61	1.04	0.47	0.78	0.65	0.70
White	43.12	44.11	40.92	40.85	38.62	43.42	35.12	37.13
Hispanic/ Latino	6.35	5.72	5.35	6.79	6.95	5.72	4.99	4.40
Asian	1.57	1.39	1.34	1.23	1.23	1.32	1.34	2.01
Native Hawaiian/ Other Pacific Islander	0.64	0.54	1.00	0.78	0.58	0.45	0.44	0.40
Two or more races	0.47	0.35	0.34	0.43	0.77	0.49	0.54	0.79
No response	25.18	24.70	27.12	27.01	28.38	29.04	48.75	46.85

Note: — indicates no students tested.

Table 3.5. 2014 ACT Aspire Writing Norm Group Demographics

	Grade (%)							
	3 (n = 25,301)	**4** (n = 26,867)	**5** (n = 27,071)	**6** (n = 29,045)	**7** (n = 30,935)	**8** (n = 32,984)	**9** (n = 15,361)	**10** (n = 13,144)
Gender								
F	47.21	45.59	46.30	46.83	46.20	46.68	43.84	42.73
M	46.01	47.74	47.12	47.15	47.07	46.36	41.22	41.68
No response	6.78	6.67	6.58	6.01	6.73	6.96	14.94	15.59
State								
AL	61.82	66.06	60.88	61.58	56.81	55.33	—	—
AR	0.07	0.07	0.06	0.08	0.05	0.06	0.10	0.13
AZ	—	—	—	—	—	0.17	0.47	0.39
CA	0.39	0.30	4.14	3.55	0.37	0.18	0.44	1.34
CO	3.36	3.10	2.79	2.06	1.87	2.18	3.19	2.91
CT	0.20	0.08	0.10	0.05	0.10	0.06	1.74	1.45
FL	0.07	0.15	—	—	—	0.11	1.38	0.73
GA	0.10	0.11	0.17	0.12	—	—	—	—
IA	0.13	0.12	0.13	0.12	0.34	0.09	0.17	1.12
IL	2.09	1.88	2.07	2.00	1.95	1.81	5.06	5.25
IN	0.83	0.49	0.75	2.20	1.92	0.79	1.35	2.58
KS	2.21	2.13	1.94	3.39	2.83	1.72	2.87	3.38
KY	6.40	6.32	5.10	3.43	3.81	2.62	5.81	0.86
LA	2.81	2.64	2.56	2.25	1.74	1.29	3.71	3.25
MA	—	—	—	—	—	—	—	0.13
MI	1.78	1.24	2.75	3.47	7.98	14.88	27.43	30.48
MN	0.44	0.41	0.46	0.63	0.44	0.42	—	0.68
MO	2.49	1.73	2.56	1.93	1.80	1.59	3.54	4.50
MS	2.99	2.77	2.12	2.52	2.09	1.76	1.87	1.38
ND	—	—	—	—	—	0.26	—	0.12
NE	—	—	—	—	—	—	0.10	0.17
NJ	—	—	—	—	—	—	0.64	—
NM	—	—	—	—	—	—	0.77	—
NV	—	—	—	—	—	—	0.09	0.10
OH	1.91	1.54	1.39	1.55	3.98	1.73	9.35	6.77

Note: — indicates no students tested.

Table 3.5. (continued)

	Grade (%)							
	3 (*n* = 25,301)	**4** (*n* = 26,867)	**5** (*n* = 27,071)	**6** (*n* = 29,045)	**7** (*n* = 30,935)	**8** (*n* = 32,984)	**9** (*n* = 15,361)	**10** (*n* = 13,144)
OK	0.98	1.08	0.94	1.34	1.27	1.14	0.19	0.14
PA	—	—	—	—	—	—	1.41	1.73
SC	4.88	4.78	4.92	4.50	4.15	4.15	2.67	2.87
SD	—	—	0.26	—	—	0.25	—	—
TN	0.53	0.52	0.64	0.67	0.61	0.67	0.87	1.71
TX	2.57	1.53	2.22	1.39	2.16	2.25	3.32	7.44
UT	0.37	0.38	0.37	0.49	1.10	0.48	—	—
VA	—	—	—	0.10	—	0.29	—	—
WI	0.57	0.57	0.69	0.56	2.61	3.62	21.00	18.00
No response	—	—	—	0.01	—	0.09	0.45	0.40
Race/ Ethnicity								
Black/African American	24.68	24.97	23.11	25.31	23.19	22.37	6.61	6.48
American Indian/Alaska Native	1.00	0.89	0.99	1.05	1.16	0.95	0.60	0.70
White	43.80	45.24	43.80	42.16	43.24	46.71	31.91	34.24
Hispanic/ Latino	6.62	6.35	5.94	7.56	7.45	6.16	4.95	5.58
Asian	1.36	1.22	1.16	1.05	1.17	1.13	0.92	1.64
Native Hawaiian/ Other Pacific Islander	0.25	0.19	0.68	0.46	0.23	0.11	0.05	0.11
Two or more races	0.42	0.33	0.33	0.39	0.77	0.48	0.59	1.01
No response	21.88	20.82	23.99	22.02	22.80	22.07	54.37	50.23

Note: — indicates no students tested.

Table 3.6. 2014 ACT Aspire English Norms: Percent of Students at or below Each Scale Score Value

Scale Score	Grade							
	3	4	5	6	7	8	9	10
400	1	1	1	1	1	1	1	1
401	1	1	1	1	1	1	1	1
402	1	1	1	1	1	1	1	1
403	1	1	1	1	1	1	1	1
404	1	1	1	1	1	1	1	1
405	2	1	1	1	1	1	1	1
406	3	1	1	1	2	1	1	1
407	5	2	1	2	2	1	2	1
408	8	2	2	2	4	2	2	2
409	12	4	2	4	4	3	3	2
410	17	6	4	5	6	3	4	3
411	24	9	5	6	7	4	5	4
412	29	13	7	8	8	6	7	5
413	36	16	10	10	10	6	8	6
414	43	21	15	13	12	8	10	7
415	50	27	18	15	14	10	13	9
416	56	31	23	18	17	11	15	10
417	63	38	27	22	20	14	16	12
418	66	44	32	26	21	17	19	14
419	72	52	37	32	26	21	22	16
420	77	55	44	35	29	22	25	19
421	81	61	48	42	33	27	28	21
422	83	69	52	46	38	31	30	22
423	88	72	60	50	43	36	33	25
424	91	77	65	56	44	39	37	28
425	93	84	68	60	50	42	40	31
426	95	85	72	66	55	47	43	33
427	95	90	79	68	60	53	47	37
428	97	92	80	73	62	55	50	40
429	97	95	86	77	67	61	54	43
430	99	96	86	81	73	67	57	47
431	99	98	91	84	75	67	61	50
432	99	98	91	87	78	73	65	53
433	99	99	95	89	83	76	68	57
434	99	99	95	92	85	78	71	60
435	100	99	97	93	87	83	74	63
436		99	98	95	91	86	77	67
437		99	98	96	92	88	80	70
438		100	99	97	94	90	82	72

Table 3.6. (continued)

Scale Score	Grade							
	3	4	5	6	7	8	9	10
439			99	98	96	92	84	75
440			99	98	97	94	87	79
441			99	99	98	95	89	81
442			100	99	99	97	91	84
443				99	99	97	93	87
444				99	99	99	94	90
445				99	99	99	96	92
446				99	99	99	97	94
447				99	99	99	98	95
448				100	99	99	99	96
449					99	99	99	98
450					100	99	99	98
451						99	99	99
452						100	99	99
453							99	99
454							99	99
455							99	99
456							100	100

Table 3.7. 2014 ACT Aspire Mathematics Norms: Percent of Students at or below Each Scale Score Value

Scale Score	Grade							
	3	4	5	6	7	8	9	10
400	1	1	1	1	1	1	1	1
401	1	1	1	1	1	1	1	1
402	1	1	1	1	1	1	1	1
403	2	1	1	1	1	1	1	1
404	3	1	1	1	1	1	1	1
405	5	1	1	1	1	1	1	1
406	8	1	1	1	2	1	1	1
407	11	3	1	2	3	1	1	1
408	16	5	3	2	5	2	1	1
409	22	6	5	3	5	3	2	2
410	32	10	8	5	8	5	2	2
411	39	17	10	8	12	7	4	4
412	50	23	16	11	13	10	6	6
413	58	34	21	16	19	13	8	6
414	69	45	31	22	25	17	11	9
415	76	55	40	29	32	24	14	12
416	84	65	50	35	39	29	18	15
417	89	74	60	42	43	34	23	19
418	94	80	64	50	50	41	28	23
419	96	85	70	57	56	46	32	26
420	98	90	76	63	61	50	37	30
421	99	93	82	70	66	55	39	32
422	99	95	87	75	71	61	44	36
423	99	96	88	79	75	65	49	40
424	99	98	91	83	79	69	53	44
425	99	98	94	86	82	72	58	48
426	99	99	95	89	86	75	61	51
427	99	99	97	91	89	79	65	54
428	99	99	98	94	91	82	69	58
429	100	99	98	95	93	84	72	61
430	100	99	99	96	94	87	75	65
431	100	99	99	97	96	88	79	68
432	100	99	99	98	97	90	82	72
433	100	99	99	99	97	92	84	75
434	100	100	99	99	98	94	88	79
435		100	99	99	98	95	90	82
436		100	99	99	99	96	92	85
437		100	99	99	99	97	93	87
438		100	99	99	99	98	95	89

Table 3.7. (continued)

Scale Score	Grade							
	3	4	5	6	7	8	9	10
439		100	99	99	99	98	96	91
440		100	99	99	99	99	97	93
441			99	99	99	99	98	95
442			99	99	99	99	98	96
443			99	99	99	99	99	97
444			99	99	99	99	99	98
445			100	99	99	99	99	99
446			100	99	99	99	99	99
447				99	99	99	99	99
448				99	100	99	99	99
449				99	100	99	99	99
450				100	100	99	99	99
451				100	100	99	99	99
452					100	100	99	99
453					100	100	99	99
454						100	99	99
455						100	99	99
456						100	99	99
457							99	100
458							99	100
459							100	100
460							100	100

Table 3.8. 2014 ACT Aspire Reading Norms: Percent of Students at or below Each Scale Score Value

Scale Score	Grade							
	3	4	5	6	7	8	9	10
400	1	1	1	1	1	1	1	1
401	1	1	1	1	1	1	1	1
402	1	1	1	1	1	1	1	1
403	2	1	1	1	1	1	1	1
404	5	1	1	1	1	1	1	1
405	8	3	2	1	1	1	1	1
406	14	6	3	2	1	1	1	1
407	23	10	5	4	3	2	2	2
408	28	14	7	6	5	2	4	3
409	36	20	10	9	5	4	4	4
410	43	25	16	13	9	5	8	6
411	48	31	21	16	13	8	11	10
412	54	36	26	18	16	10	15	13
413	61	45	31	22	21	13	16	14
414	66	51	35	26	25	17	21	17
415	72	57	40	31	31	20	26	21
416	77	63	49	35	33	23	30	24
417	82	68	55	40	38	27	34	28
418	87	75	61	46	44	32	39	31
419	91	80	67	50	48	36	43	35
420	94	84	73	59	53	41	47	39
421	95	86	78	61	59	47	52	43
422	96	90	80	67	65	49	57	47
423	98	93	84	73	71	55	61	51
424	99	96	86	78	76	58	62	52
425	99	97	91	83	81	64	66	57
426	99	98	94	87	86	70	71	62
427	99	99	94	89	90	76	75	66
428	99	99	97	93	93	82	80	72
429	100	99	98	96	96	87	84	77
430		99	99	96	96	89	89	82
431		100	99	98	98	91	89	83
432			99	99	99	95	92	87
433			99	99	99	97	95	92
434			100	99	99	98	97	95
435				99	99	99	99	98
436				100	99	99	99	98
437					99	99	99	99
438					100	99	99	99

Table 3.8. (continued)

Scale Score	Grade							
	3	4	5	6	7	8	9	10
439						99	99	99
440						100	99	99
441							99	99
442							100	100

Table 3.9. 2014 ACT Aspire Science Norms: Percent of Students at or below Each Scale Score Value

Scale Score	Grade							
	3	4	5	6	7	8	9	10
400	1	1	1	1	1	1	1	1
401	1	1	1	1	1	1	1	1
402	1	1	1	1	1	1	1	1
403	2	1	1	1	1	1	1	1
404	4	2	1	1	1	1	1	1
405	6	3	2	2	1	1	1	1
406	10	6	3	2	2	1	1	1
407	15	8	4	4	3	1	1	1
408	22	12	6	5	5	2	2	2
409	28	15	8	8	7	3	2	2
410	34	18	11	10	10	5	4	4
411	39	22	14	13	14	7	4	4
412	46	26	18	17	18	9	7	6
413	51	31	22	22	21	12	7	7
414	58	37	26	25	26	15	11	9
415	62	42	30	29	28	19	16	14
416	66	47	34	34	32	23	17	15
417	71	54	39	38	38	27	22	19
418	76	60	45	42	42	28	24	21
419	81	65	51	47	46	33	29	25
420	84	70	57	51	50	37	35	29
421	88	75	63	57	56	42	40	34
422	91	79	68	62	60	46	45	38
423	94	84	74	67	64	51	46	39
424	95	87	78	72	67	57	50	42
425	97	90	83	77	72	62	54	46
426	98	93	89	81	76	66	59	50
427	99	95	92	85	80	70	63	53
428	99	97	94	89	83	75	67	57
429	99	98	96	91	86	79	70	61
430	99	99	97	93	88	82	76	66
431	99	99	98	95	91	85	79	69
432	99	99	99	96	94	88	82	73
433	100	99	99	98	96	90	84	76
434		99	99	99	98	92	87	79
435		99	99	99	99	95	91	83
436		100	99	99	99	96	92	86
437			99	99	99	97	94	88
438			100	99	99	98	96	91

Table 3.9. (continued)

Scale Score	Grade							
	3	4	5	6	7	8	9	10
439				99	99	99	97	94
440				100	99	99	98	95
441					99	99	99	97
442					99	99	99	98
443					100	99	99	99
444						99	99	99
445						99	99	99
446						100	99	99
447							99	99
448							99	99
449							100	100

Table 3.10. 2014 ACT Aspire Writing Norms: Percent of Students at or below Each Scale Score Value

Scale Score	Grade							
	3	4	5	6	7	8	9	EHS
408	1	1	2	1	1	2	1	1
409	1	1	2	2	1	2	1	1
410	4	4	5	2	5	4	5	4
411	4	4	5	2	5	5	5	4
412	7	5	7	3	9	7	7	6
413	7	5	7	3	9	7	7	6
414	11	7	8	7	11	9	9	7
415	11	7	8	7	12	9	9	7
416	20	14	18	14	19	19	13	10
417	20	14	18	15	19	19	13	10
418	33	27	27	18	29	31	26	20
419	33	29	27	19	30	31	26	20
420	51	40	35	22	40	39	31	24
421	51	40	35	22	41	39	31	24
422	63	51	40	35	49	44	35	27
423	63	51	40	42	49	44	35	27
424	74	65	58	51	65	61	47	37
425	75	65	58	51	65	61	47	37
426	83	80	73	58	73	74	65	55
427	84	81	74	58	74	74	65	55
428	89	87	78	61	81	81	71	61
429	89	87	79	61	82	81	71	61
430	94	92	82	64	86	84	76	66
431	94	92	82	80	87	84	76	66
432	97	97	93	87	95	95	88	81
433	97	97	93	89	95	95	88	81
434	98	98	96	91	96	97	95	90
435	98	98	96	91	96	97	95	90
436	99	99	97	92	98	98	96	92
437	99	99	97	93	98	98	96	92
438	99	99	98	94	98	98	97	94
439	99	99	98	94	98	98	97	94
440	100	100	100	99	99	99	99	98
441				99	99	99	99	98
442				99	99	99	99	99
443				99	99	99	99	99
444				99	99	99	99	99
445				99	99	99	99	99

Table 3.10. (continued)

Scale Score	Grade							
	3	4	5	6	7	8	9	EHS
446				99	99	99	99	99
447				99	99	99	99	99
448				100	100	100	100	100

CHAPTER 4

EPAS® to ACT Aspire Concordance

A concordance study was conducted between the Educational Planning and Assessment System (EPAS®) 1–36 scale, which consists of The ACT (grades 11–12) and two legacy tests, ACT Explore® (grades 8–9) and ACT Plan® (grade 10), and the three-digit ACT Aspire scale. The concordance study established a direct link between scores on EPAS and ACT Aspire. This link was used to facilitate a smooth transition to ACT Aspire for users of ACT Explore and ACT Plan, and to establish the ACT Readiness Benchmarks for grades 3–11 (described in chapter 5), and to establish the Progress toward Career Readiness indicator (chapter 7).

EPAS is an integrated series of paper-administered, curriculum-based tests of educational development with selected-response items typically taken by students from grade 8 through high school. ACT Aspire, on the other hand, is offered on paper or online; includes selected-response, constructed-response, and technology-enhanced item types; and is a vertically-articulated, benchmarked, and standards-based system of assessments that can be taken by students from grades 3 through early high school. The grade 8 through early high school tests in EPAS and ACT Aspire are intended to measure similar constructs but differ in test specifications, which is a circumstance where concordances are an applicable type of scale alignment (Holland and Dorans 2006).

Method

The concordance analysis involved relating scores from ACT Explore, ACT Plan, and The ACT to ACT Aspire using percentile ranks, where concorded scores are defined as those having the same percentage of students at or below a given score with respect to the group of students used in the study. The collection of concorded scores is referred to as a concordance table, which is useful for determining the cut scores on one test that result in approximately the same *proportion* of students identified by the other test, although not necessarily the same students.

The ACT Aspire scale scores for students in the concordance analysis were from the spring 2013 special studies in which participating students took ACT Aspire English, Mathematics, Reading, or Science tests plus historical ACT Explore, ACT Plan, or The ACT assessments from the 2012–2013 academic year. Note that grade 11 students took The ACT plus a separately timed section of constructed-response items designed as an add-on to The ACT (see appendix A). The constructed-response portion was combined with an operationally administered form of The ACT (selected response only) to obtain scores on the ACT Aspire scale. This version of The ACT is referred to as The *ACT with constructed-response tests*. The final analysis sample included students from grade 7 and above who took one of the EPAS tests (ACT Explore, ACT Plan, or The ACT) and the ACT Aspire test in the same subject and grade level (which included The ACT with constructed response for grade 11 students).

Since both EPAS and ACT Aspire are vertically scaled assessments covering different grade spans, the concordance was established to link between two scales (i.e., the 1–36 EPAS scale to the ACT Aspire 400+ scale), rather than between two tests (e.g., between ACT Explore and ACT Aspire Grade 8 test).

The concordance relationship between EPAS and ACT Aspire scales was estimated using the equipercentile method described by Kolen and Brennan (2014). This method uses the percentile ranks (i.e., the proportion of scores at or below each score) to define the relationship between the two scales. For a particular EPAS score, the corresponding ACT Aspire score is the one that has the same percentile rank.

Evaluation of Results

The derived concordance table was applied to the existing datasets with students' ACT Explore and ACT Plan scores. The purpose of the evaluation was to examine distributions of the concorded ACT Aspire scale scores that should be similar to those of the original EPAS scale scores. Two evaluation samples were used. One sample was composed of cross-sectional data in the academic year 2013–2014 with the cohort of grade 8 students who took ACT Explore and the cohort of grade 10

students who took ACT Plan. The other sample was a longitudinal data set including the cohort of grade 8 students in the academic year 2010–2011 who took ACT Explore in that year and took ACT Plan two years later when they were in grade 10. The following properties were examined:

1. Using the cross-sectional data, compare distributions of ACT Explore/ACT Plan scores against the distributions of concorded ACT Aspire scale scores. The purpose of this analysis was to verify that ACT Explore/ACT Plan and concorded ACT Aspire score distributions did not appear to differ.
2. Using the longitudinal data, create box plots of ACT Plan scores conditional on each ACT Explore score point, and create the same set of plots based on the concorded ACT Aspire scale scores. The purpose of this analysis was to examine whether the relationship between grade 8 and grade 10 scores of the same cohort stayed the same when the concorded ACT Aspire scores were used.
3. Using the longitudinal data, compute effect sizes from grade 8 to grade 10, based on the EPAS scale score and the concorded ACT Aspire scale scores. The effect size was computed as the scale score difference between the two grades divided by the square root of the average variances. The purpose of this analysis was to examine whether the magnitude of growth measured by effect size was similar when the concordance was applied.

Results

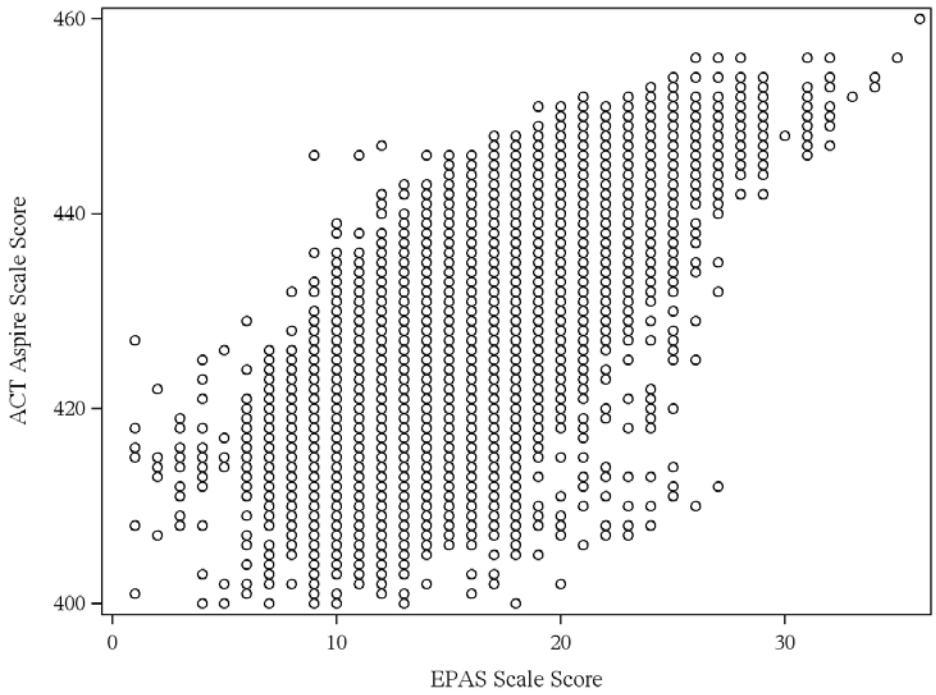
Table 4.1 presents the *n*-counts of the analysis sample, the correlation between the EPAS and ACT Aspire scale scores, and descriptive statistics of scores on both scales. As shown, a sample of over 16,000 students who had both EPAS and ACT Aspire scores was used to conduct the concordance analysis for four subject areas: English, mathematics, reading and science. Correlations between the two scale scores are higher in English and mathematics than in reading and science. The scatter plots between the two sets of scores, shown in figure 4.1, indicate a linear relationship.

Table 4.1. Descriptive Statistics of the Sample Used in the Concordance Analysis

Subject	N	Correlation between EPAS and ACT Aspire	EPAS				ACT Aspire			
			Mean	SD	Min	Max	Mean	SD	Min	Max
English	18,204	.81	16.04	4.87	1	36	427.68	10.24	400	460
Math	17,787	.80	16.50	4.37	1	36	423.44	8.62	401	457
Reading	16,922	.70	16.02	4.72	1	36	420.74	7.57	401	446
Science	17,102	.74	17.62	3.87	1	36	423.06	8.32	400	449

Figure 4.1. Scatter plots between ACT Aspire scale score and EPAS scale scores

English



Mathematics

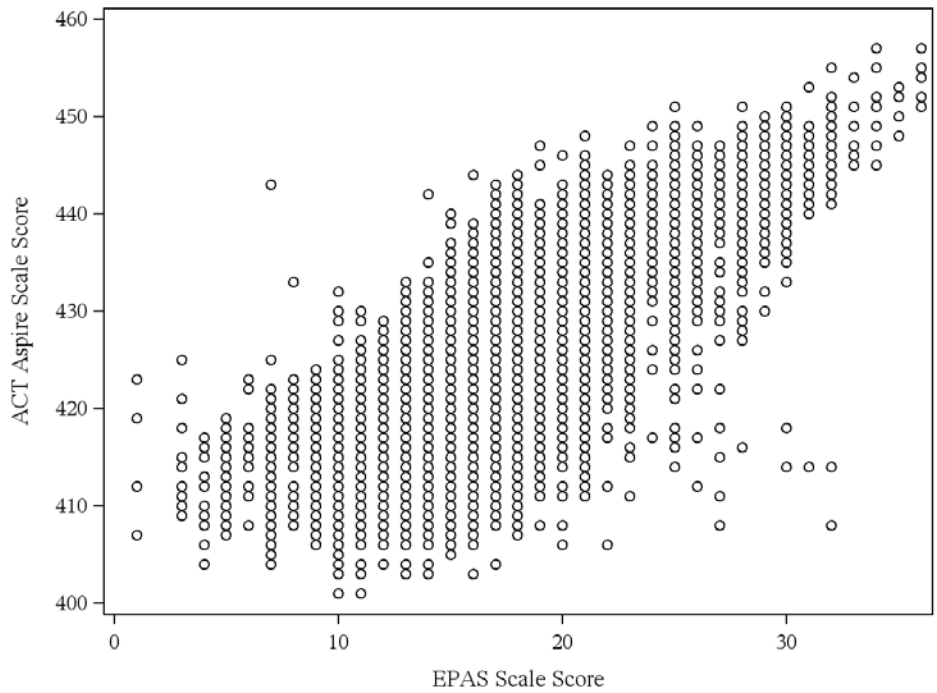
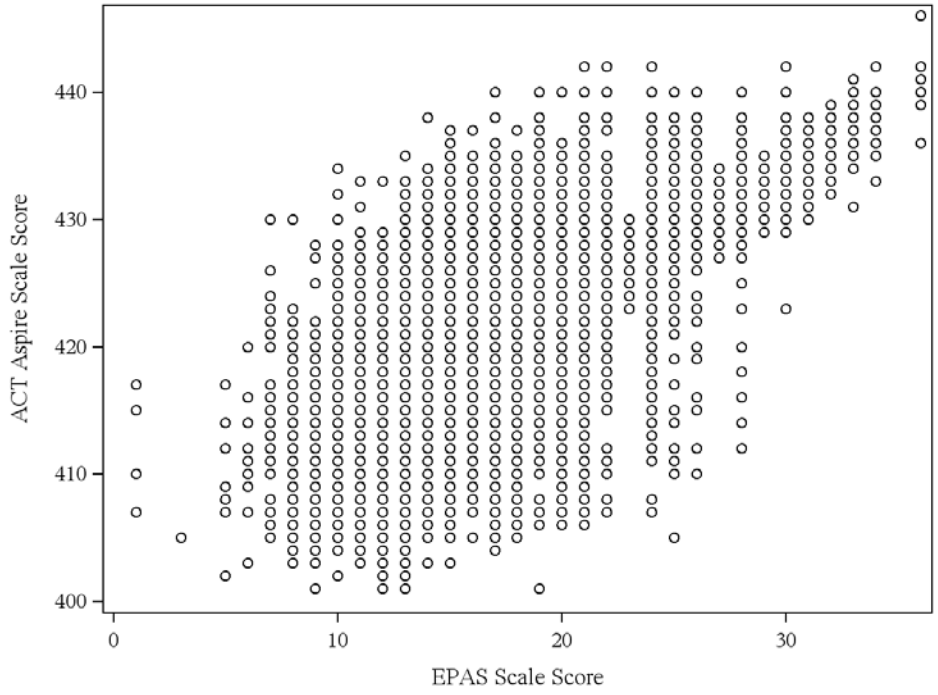
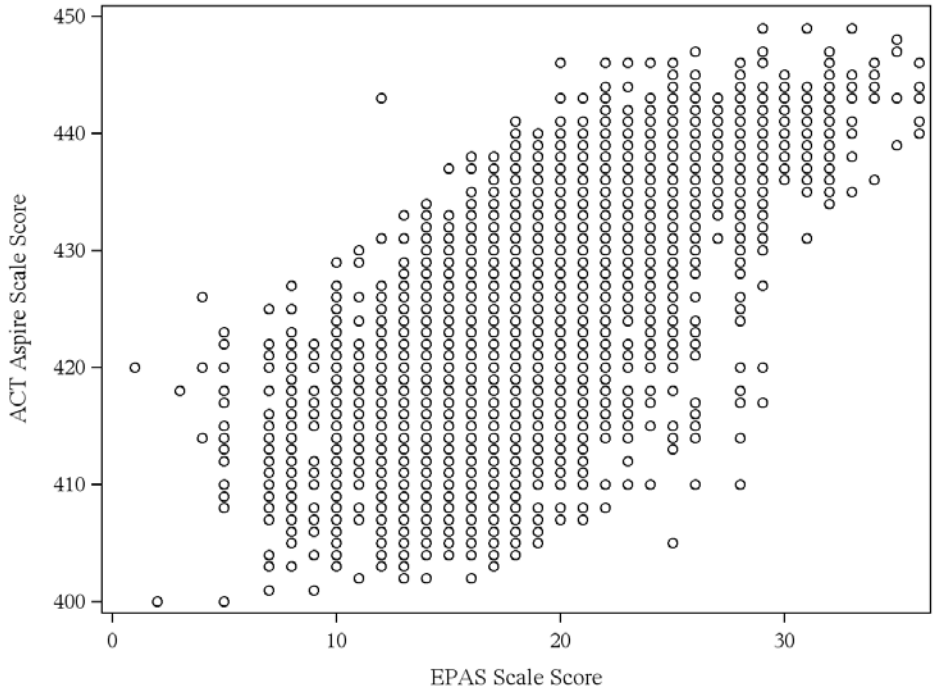


Figure 4.1. (continued)

Reading



Science



Evaluating Results

Descriptive statistics of ACT Explore and the concorded ACT Aspire scores in the cross-sectional evaluation sample are shown in table 4.2, and the scores for ACT Plan are shown in table 4.3. Figure 4.2 presents the distributions of ACT Explore and concorded ACT Aspire scores, and figure 4.3 presents similar information for ACT Plan scores. It can be seen that the concorded score distributions match those of the original scores.

For each pair of panels in figure 4.4, the top panel presents the box plots of ACT Plan scores conditional on ACT Explore scores using the longitudinal evaluation sample. The bottom panel presents the plots using the concorded ACT Aspire score. As shown, the relationship between students' grade 10 and grade 8 performance is highly similar when the concordance was applied.

Table 4.2. Descriptive Statistics of the ACT Explore and Concorded ACT Aspire Scores in Cross-Sectional Evaluation Sample

		ACT Explore				Concorded ACT Aspire			
Subject	**N**	**Mean**	**STD**	**Min**	**Max**	**Mean**	**STD**	**Min**	**Max**
English	1,034,067	14.40	4.37	1	25	424.17	9.64	400	445
Math	1,033,788	15.40	3.73	1	25	421.33	7.71	400	439
Reading	1,032,226	14.56	3.99	1	25	418.23	7.05	400	434
Science	1,030,623	16.54	3.33	1	25	420.78	7.51	400	438

Table 4.3. Descriptive Statistics of ACT Plan and Concorded ACT Aspire Scores in Cross-Sectional Evaluation Sample

		ACT Plan				Concorded ACT Aspire			
Subject	**N**	**Mean**	**STD**	**Min**	**Max**	**Mean**	**STD**	**Min**	**Max**
English	1,251,757	16.94	4.72	1	32	429.59	9.72	400	453
Math	1,251,809	17.74	4.85	1	32	425.83	9.25	400	452
Reading	1,250,597	17.15	4.66	1	30	422.66	7.21	400	437
Science	1,250,022	18.58	4.13	1	32	425.01	8.53	400	447

Figure 4.2. Distributions of ACT Explore scale scores and concorded ACT Aspire scale scores

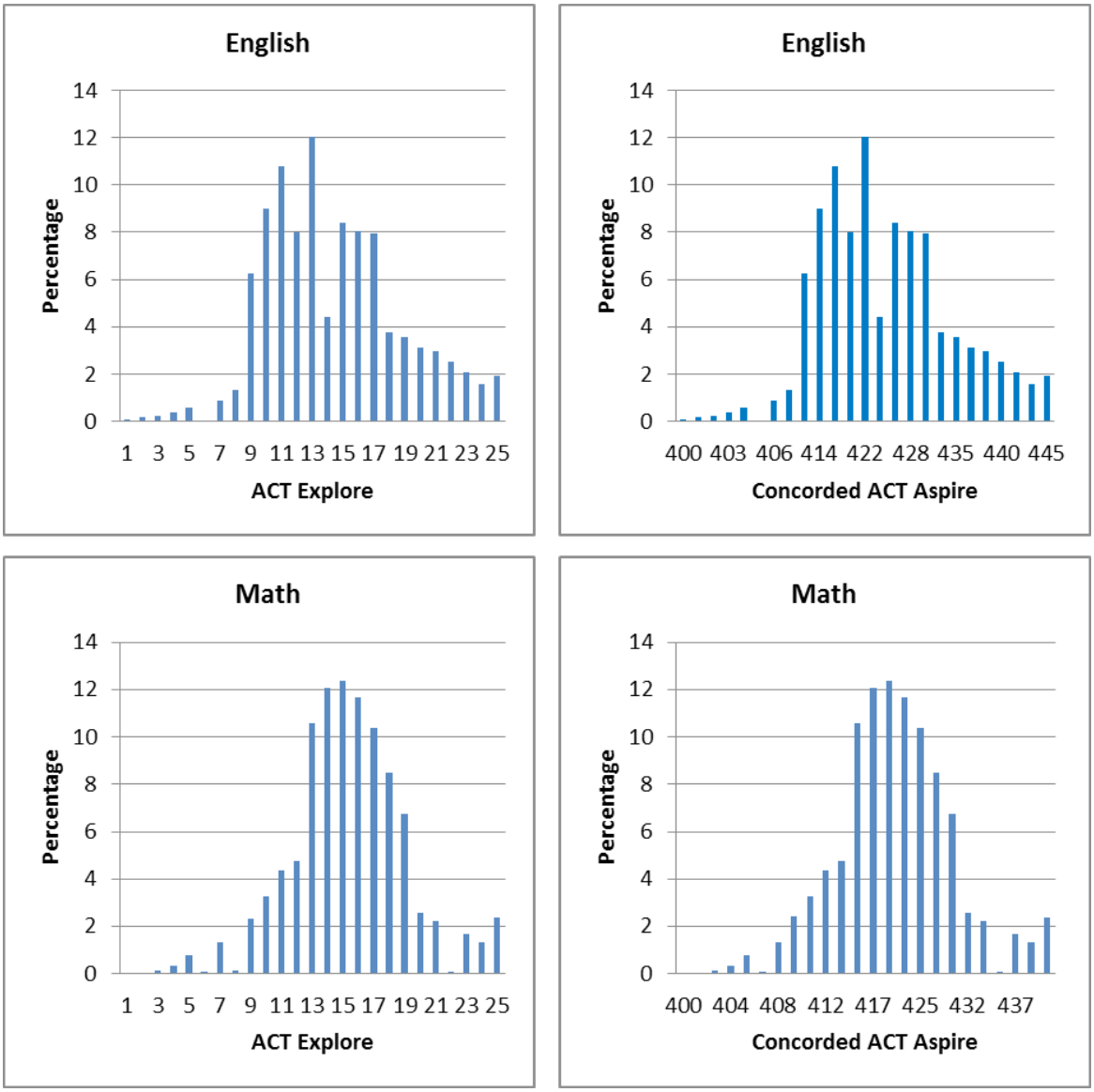


Figure 4.2. (continued)

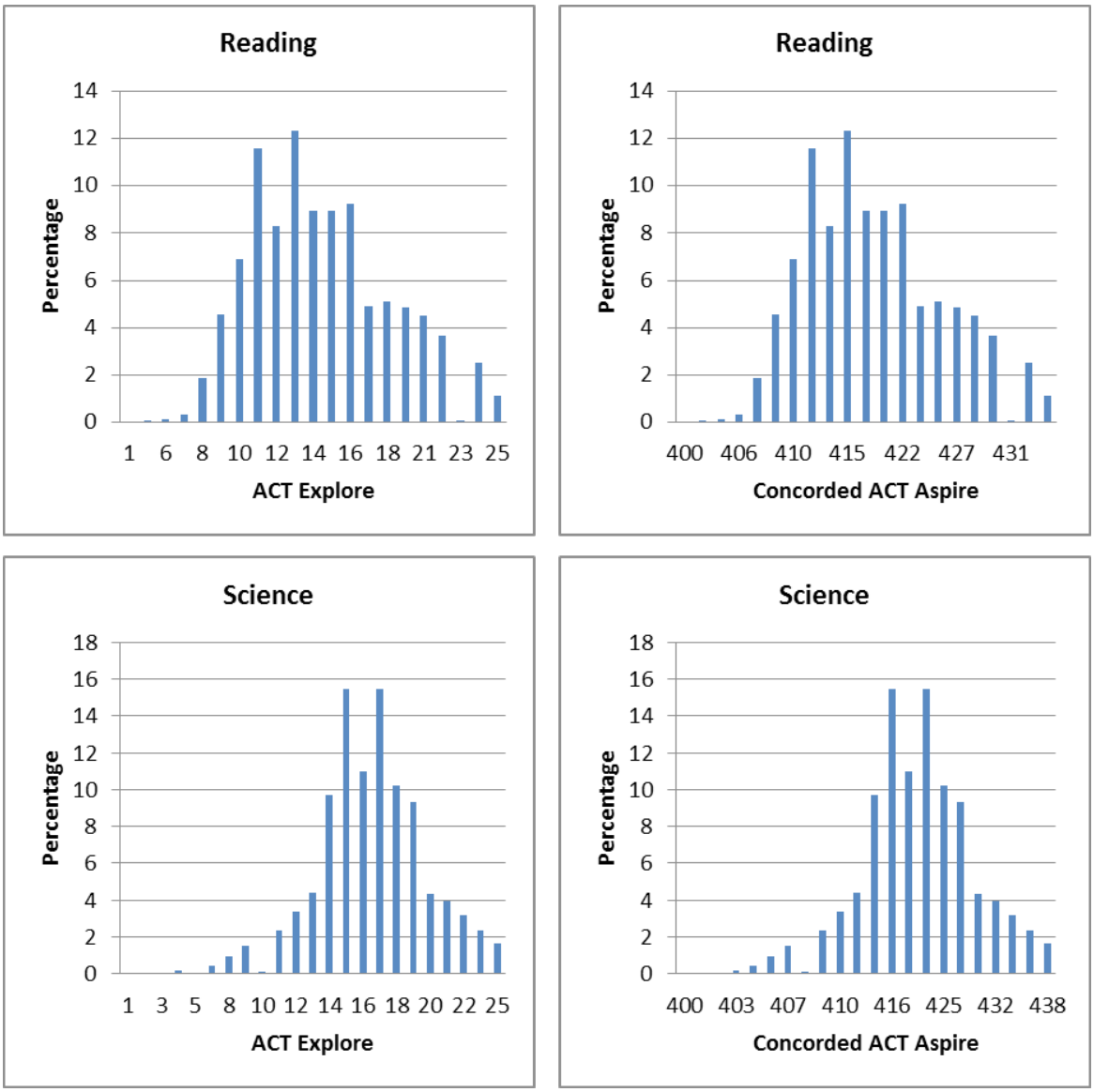


Figure 4.3. Distributions of ACT Plan scale scores and concorded ACT Aspire scale scores

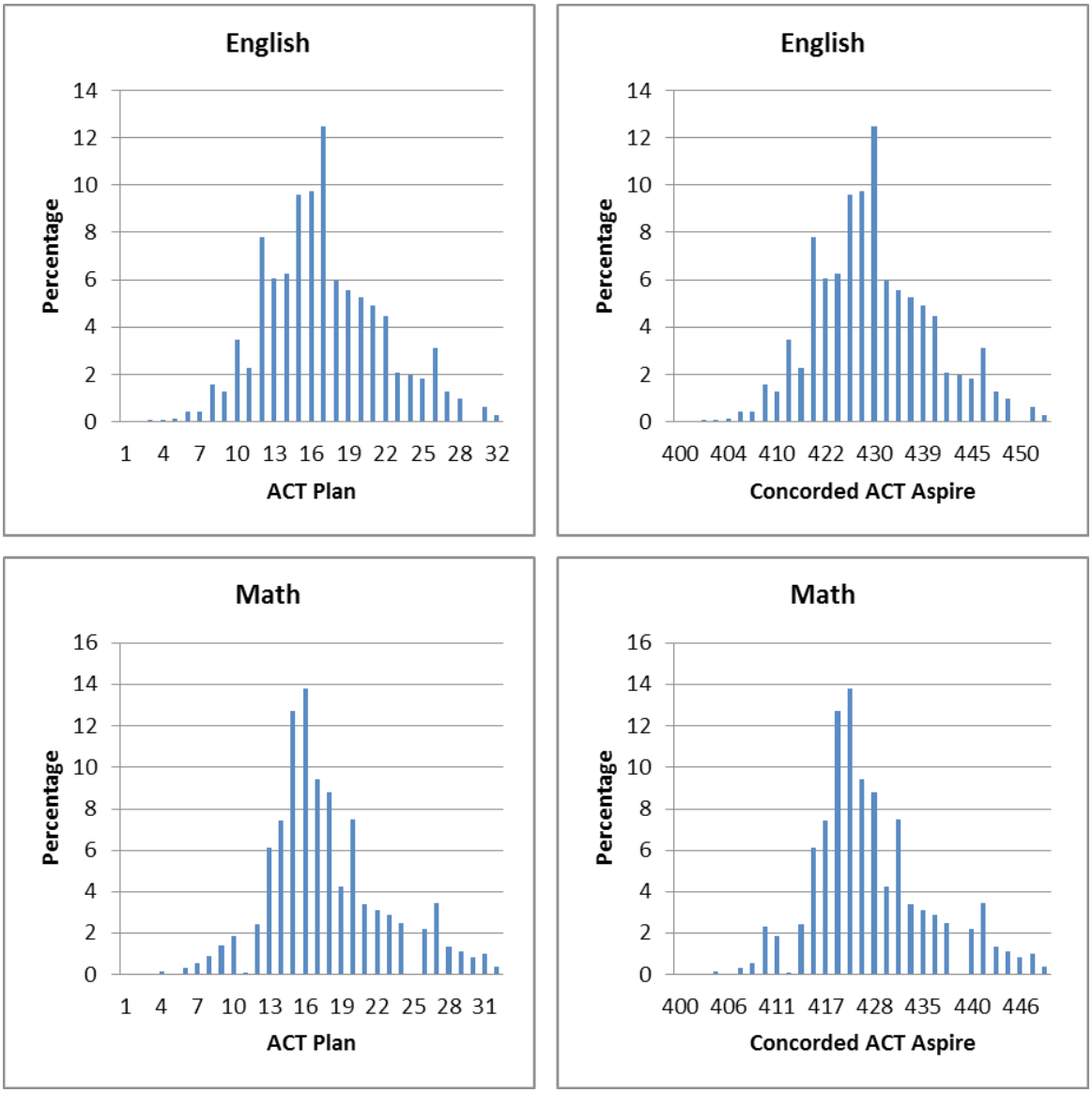


Figure 4.3. (continued)

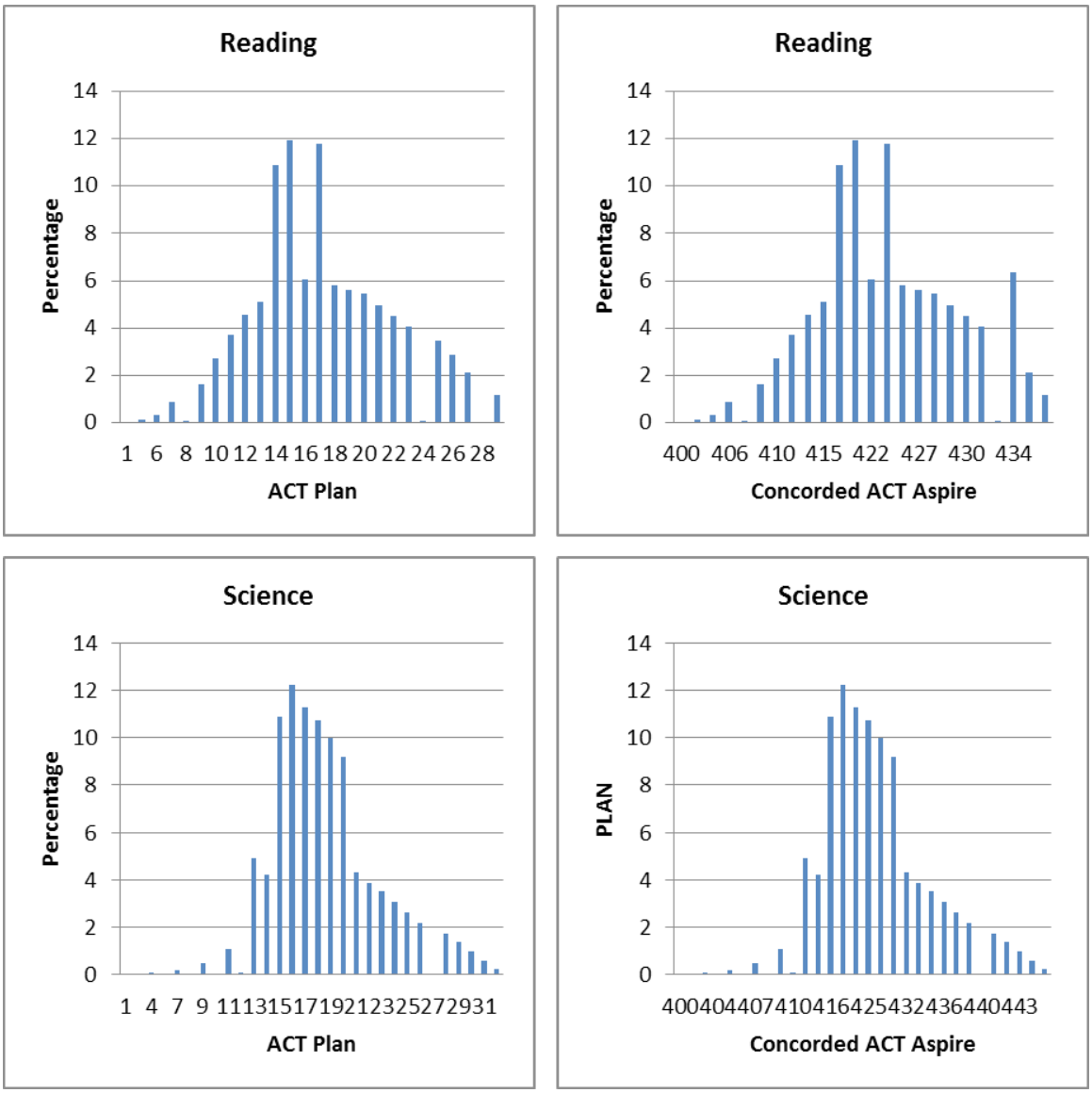
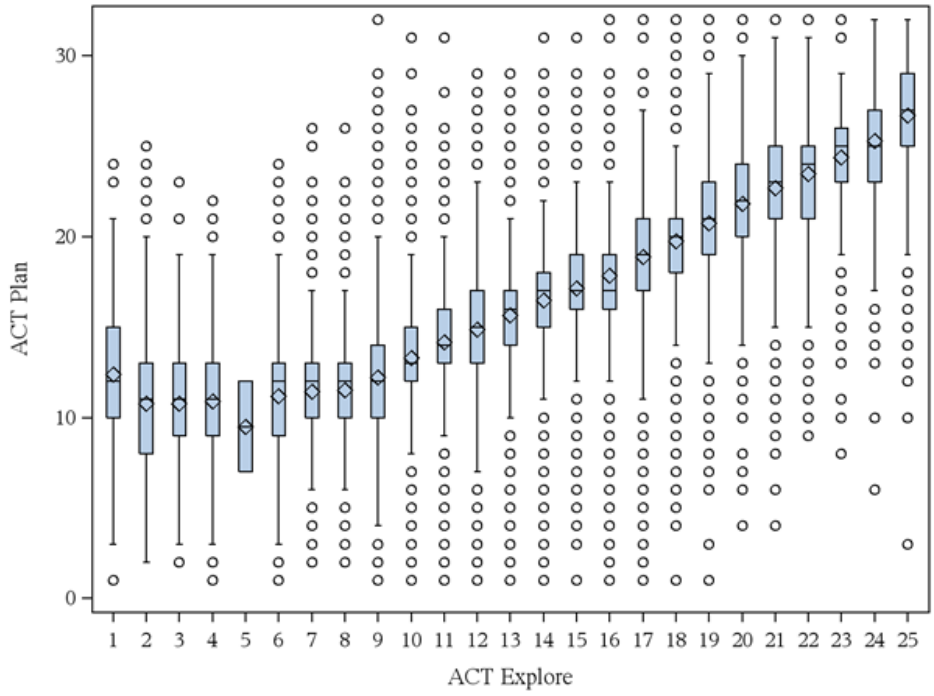


Figure 4.4. Box plots of ACT Plan (or concorded ACT Aspire scale scores) conditional on ACT Explore (or concorded ACT Aspire Scores)

English—EPAS Scale



English—Concorded ACT Aspire Scale

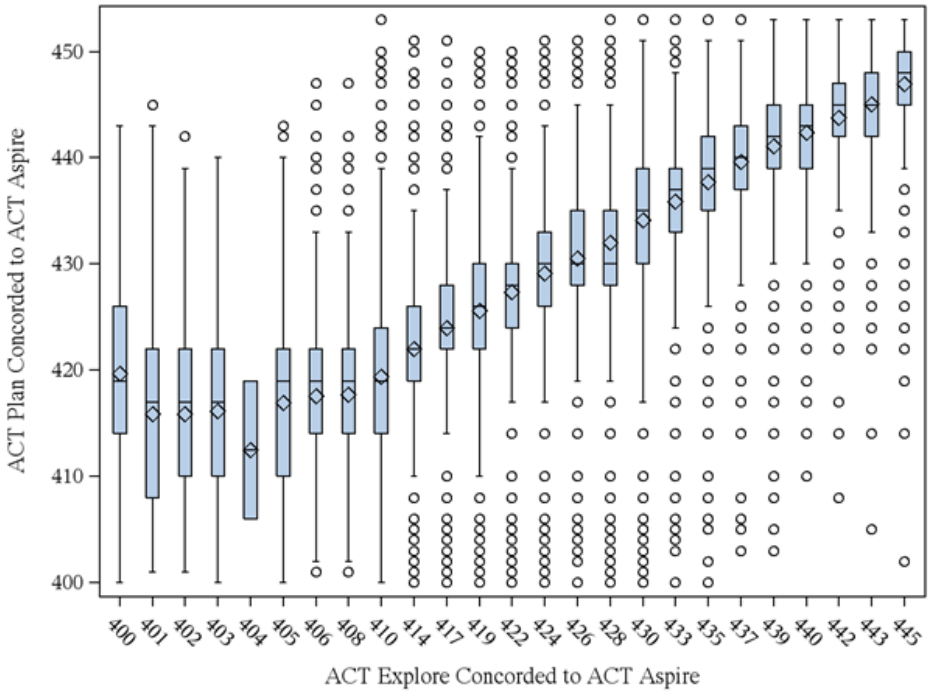
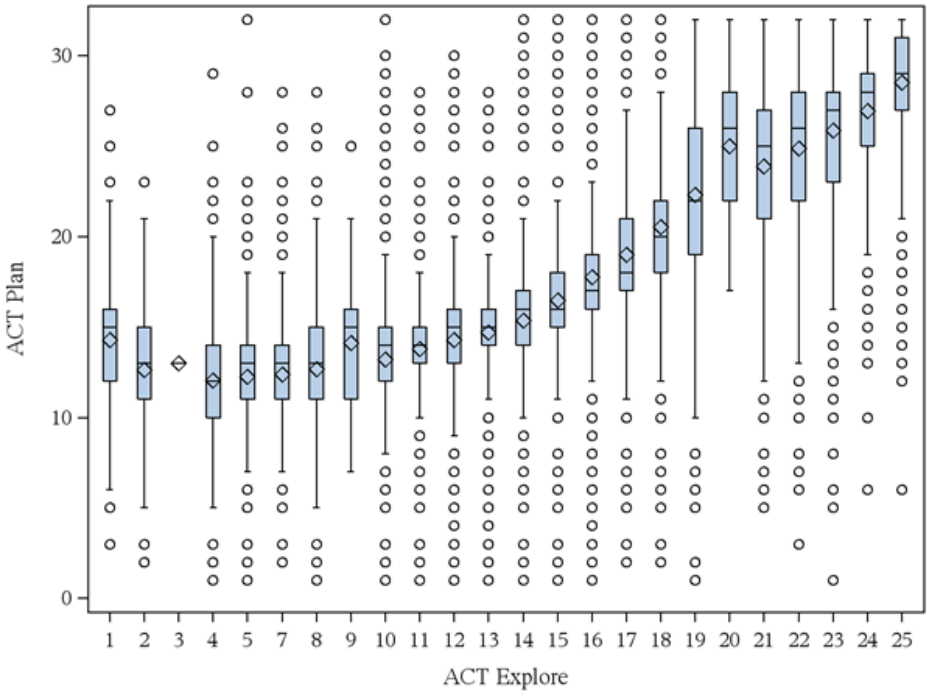


Figure 4.4. (continued)

Mathematics—EPAS Scale



Mathematics—Concorded ACT Aspire Scale

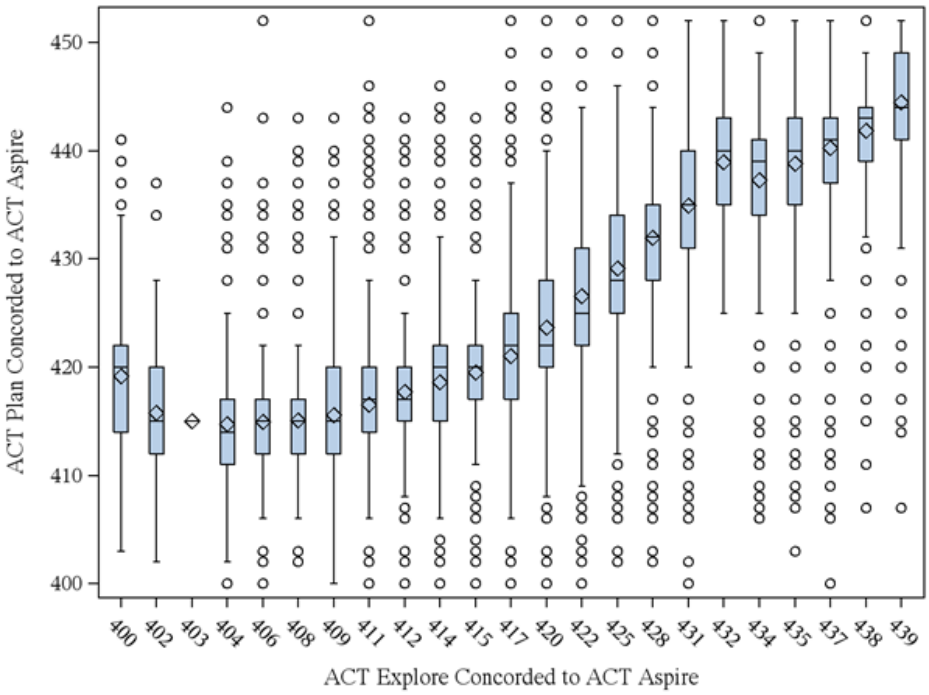
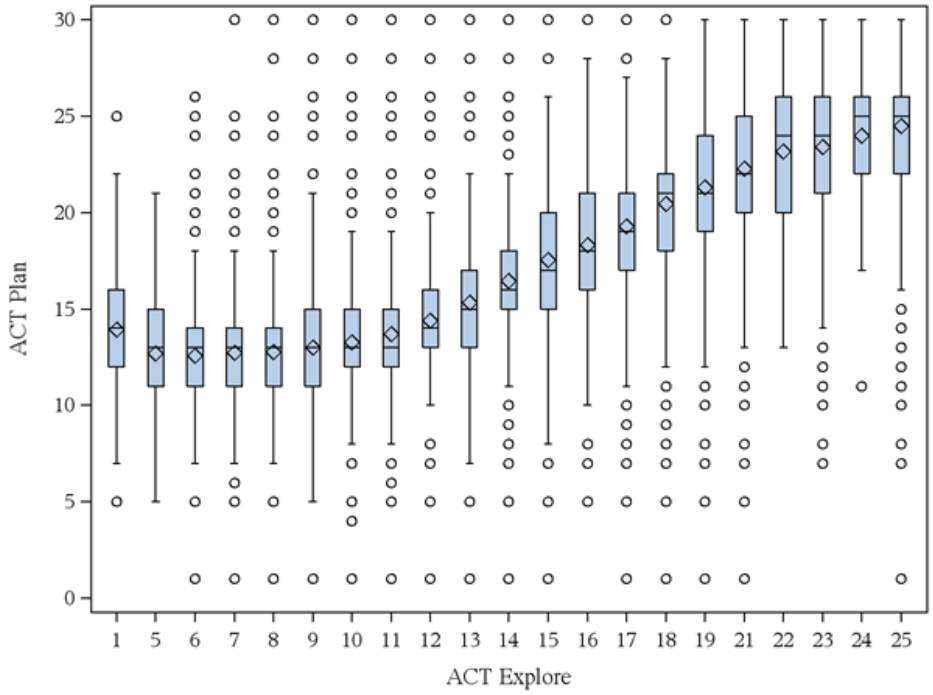


Figure 4.4. (continued)

Reading—EPAS Scale



Reading—Concorded ACT Aspire Scale

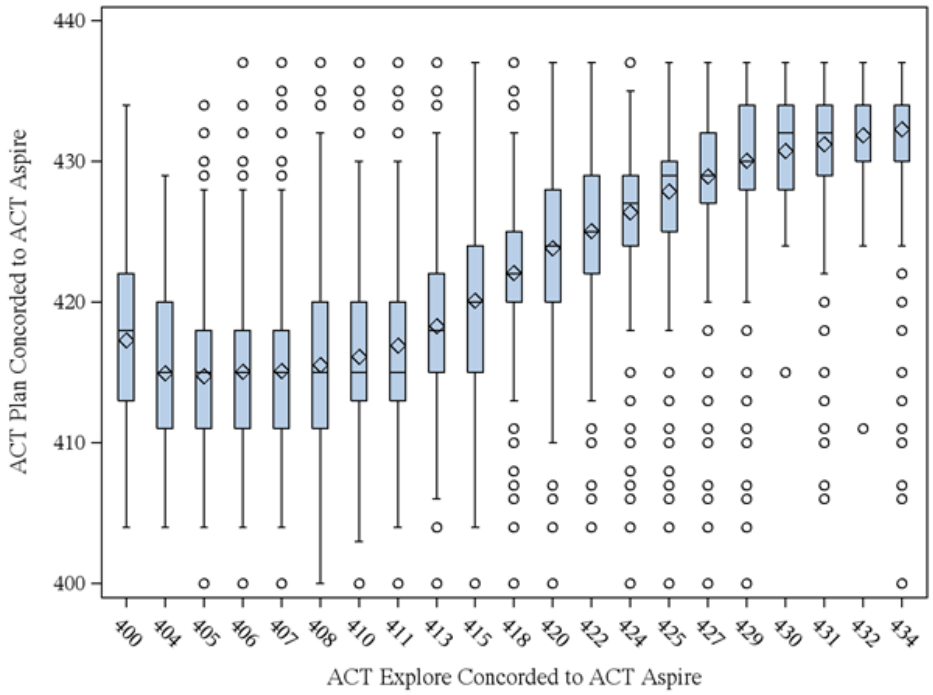
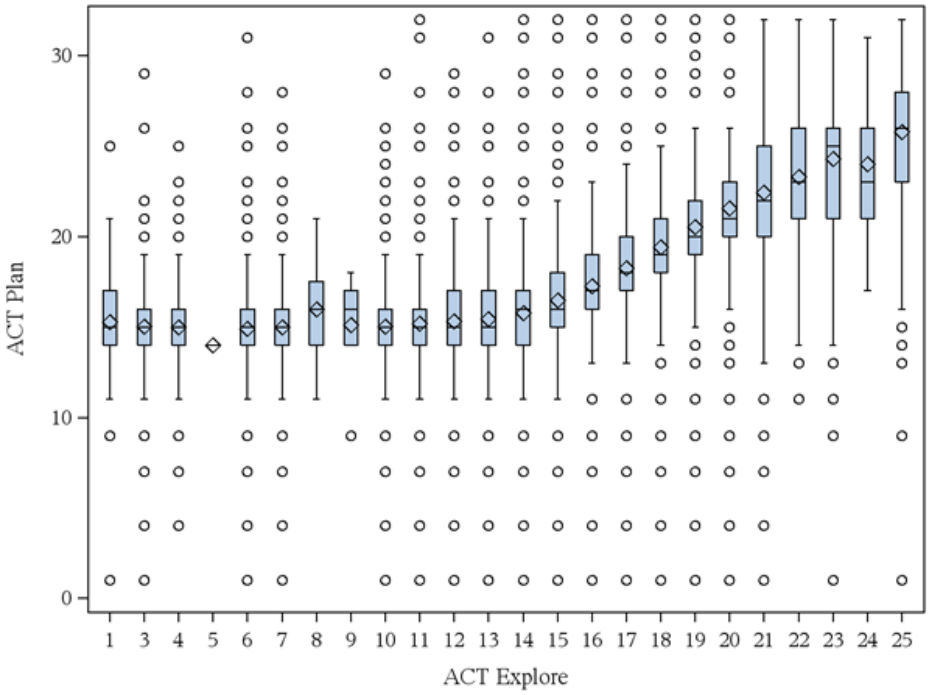
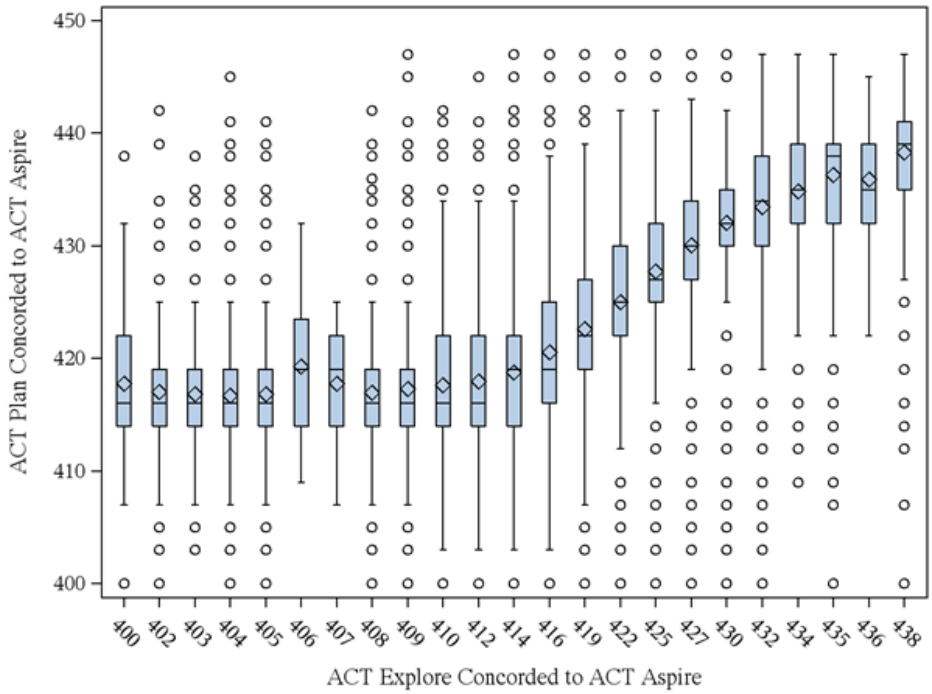


Figure 4.4. (continued)

Science—EPAS Scale



Science—ACT Aspire Concorded Scale



Descriptive statistics of the ACT Explore and ACT Plan scale scores for the longitudinal sample are presented in table 4.4. The corresponding concorded ACT Aspire scores are presented in table 4.5. Effect sizes, computed from both scales shown in the last column for each table, are very similar between the EPAS scale and the concorded ACT Aspire scale.

The final derived EPAS to ACT Aspire concordance table is presented in appendix B.

Summary

The primary purpose of the concordance was to facilitate the transition from ACT Explore and ACT Plan assessments to ACT Aspire. While the EPAS to ACT Aspire concordance tables provide a link between the two scales, the concordance should be used cautiously. For example, the sample of students included in the concordance analysis may not be representative of all EPAS or ACT Aspire test-takers. In addition, population dependence often results when the two tests that are linked do not measure the same construct. It is generally inappropriate to use the concorded scores to estimate individual student performance on the EPAS/ACT Aspire tests because scores resulting from concordance are not viewed as interchangeable with actual scores. It is preferable to use actual scores from the EPAS or ACT Aspire when possible.

Table 4.4. Descriptive Statistics of ACT Explore Grade 8 and ACT Plan Grade 10 in the Longitudinal Data

		Grade 8—ACT Explore				Grade 10—ACT Plan				
Subject	**N**	**Mean**	**STD**	**Min**	**Max**	**Mean**	**STD**	**Min**	**Max**	**Effect Size**
English	572,302	14.66	4.17	1	25	17.10	4.42	1	32	.57
Math	572,051	15.63	3.46	1	25	17.84	4.57	1	32	.55
Reading	571,328	14.80	3.91	1	25	17.19	4.48	1	30	.57
Science	570,063	16.80	3.31	1	25	18.52	3.77	1	32	.48

Table 4.5. Descriptive Statistics of Concorded ACT Aspire in the Longitudinal Data

		Grade 8—Concorded ACT Aspire				Grade 10—Concorded ACT Aspire				
Subject	**N**	**Mean**	**STD**	**Min**	**Max**	**Mean**	**STD**	**Min**	**Max**	**Effect Size**
English	572,302	424.79	9.31	400	445	430.05	9.08	400	453	.57
Math	572,051	421.85	7.24	400	439	426.12	8.76	400	452	.53
Reading	571,328	418.73	6.93	400	434	422.69	6.96	400	437	.57
Science	570,063	421.39	7.54	400	438	425.10	8.04	400	447	.48

CHAPTER 5

ACT Readiness Benchmarks

Introduction

The ACT College Readiness Benchmarks are cornerstones of The ACT and the legacy assessments ACT Explore (grades 8 and 9) and ACT Plan (grade 10), which together form EPAS. The ACT Benchmarks were established to reflect college and career readiness. The Benchmark on each of four subject tests of The ACT (English, Mathematics, Reading, and Science) are the score on the 1–36 EPAS scale at which students have a 50% probability of attaining a grade of B or higher or about a 75% chance of obtaining a C or higher in selected credit-bearing first-year college courses (for additional details, see ACT 2007b).

The ACT Readiness Benchmarks used with ACT Aspire were created to be aligned with the ACT College Readiness Benchmarks used with EPAS. Similar to EPAS, each ACT Aspire grade and subject has its own ACT Readiness Benchmark. Students at or above the benchmark are on target to meet the corresponding ACT College Readiness Benchmarks in grade 11.

ACT Readiness Benchmarks for English, Mathematics, Reading, and Science

For English, mathematics, reading, and science, the ACT Readiness Benchmarks for grades 8 through 10 were derived using the EPAS to ACT Aspire concordance tables (see chapter 4 or appendix B). Benchmarks for grades 3–7 were created using a backmapping procedure.

Grades 8–10

A concordance between the EPAS 1–36 scale and the ACT Aspire three-digit scale was used to establish ACT Readiness Benchmarks for grades 8–10 using ACT College Readiness Benchmarks that had already been established for EPAS. The concordance study included students taking both EPAS and ACT Aspire. The concordance between EPAS scale scores and ACT Aspire scale scores was obtained using equipercentile linking. The ACT College Readiness Benchmarks for grades 8, 9, and 10 were used to identify the corresponding concorded ACT Aspire scores, which were then defined as the ACT Readiness Benchmarks for grades 8–10. Note that the concorded benchmarks derived for grades 8–10 may be updated as more longitudinal data become available.

The ACT College Readiness Benchmarks used in this analysis were obtained by using an alternate set of benchmarks than those typically reported for ACT Explore and ACT Plan. This alternate set of benchmarks was based on students tested in spring, while EPAS benchmarks typically used for grades 8–10 are based on students tested in fall. ACT Aspire is administered in both fall and spring, but it is expected that benchmarks would differ slightly within the same grade due to additional instruction and academic achievement, which occurs throughout the school year. Spring benchmarks were used because they anchored the ACT Readiness Benchmarks to spring levels of educational development, near the end of a particular grade level when ACT Aspire was anticipated to be most commonly administered. Therefore, the ACT Readiness Benchmarks used for ACT Aspire are based on students tested in spring and reflect performance levels of students toward the end of an academic year. This should be kept in mind when interpreting results from fall testing. Table 5.1 presents the ACT College Readiness Benchmarks on the 1–36 scale for grades 8, 9, and 10 used in this study and the corresponding ACT Readiness Benchmarks, which are the concorded ACT Aspire scores obtained from the concordance.

Table 5.1. ACT College Readiness Benchmark and ACT Readiness Benchmark for Grades 8–10

Subject	Grade	ACT College Readiness Benchmark (Spring)	ACT Readiness Benchmarks (Concorded ACT Aspire Scores)
English	8	13	422
	9	15	426
	10	16	428
Math	8	17	425
	9	18	428
	10	20	432
Reading	8	17	424
	9	18	425
	10	20	428
Science	8	19	427
	9	20	430
	10	21	432

Agreement rates for classifying students at or above both sets of benchmarks were compared using data from the concordance study. There was agreement if students were classified into the same category using both the EPAS benchmarks and the ACT Aspire benchmarks for their particular grade. As shown in table 5.2, the agreement rates for all subjects were at or above 80%.

Grades 3–7

ACT Readiness Benchmarks for grades 3–7 in English, mathematics, reading, and science were backmapped from the grade 8 ACT Readiness Benchmark described above. Backmapping used a *z*-score approach, which involved identifying the ACT Aspire scores at a standardized score (*z*-score) that corresponded to the *z*-score for the grade 8 ACT Readiness Benchmark. Spring 2013 ACT Aspire special study data were used to create benchmarks for grades 3–7. Table 5.3 shows the backmapped ACT Readiness Benchmarks for grades 3–7. Note that the backmapped benchmarks derived for grades 3–7 will be reviewed and may be updated as more longitudinal data become available.

Table 5.2. Classification Agreement Rate between ACT Aspire and EPAS Benchmarks

	ACT Aspire Classification Agreement Rate	
	ACT Explore	**ACT Plan**
English	81%	80%
Math	82%	85%
Reading	81%	81%
Science	83%	84%

Table 5.3. ACT Readiness Benchmarks, Grades 3–7

Grade	English	Math	Reading	Science
3	413	413	415	418
4	417	416	417	420
5	419	418	420	422
6	420	420	421	423
7	421	422	423	425

ACT Readiness Benchmark for Writing

The ACT Readiness Benchmark for writing was set as the scale score corresponding to raw scores at a cut determined by ACT content experts, based on their knowledge, expertise, and experience with writing content and their development of the rubrics used to score writing prompts. A trait score of 4 formed the basis for the writing cut, which corresponds to performance consistent with grade-level expectations on the rubric. Recall that writing scores consist of four traits, each scored on a rubric with scores from 1 to 5 (grades 3–5) or 1 to 6 (grades 6–10). Content experts determined that a trait score of 4 was at the expected readiness level for each grade level. Across the four traits, if a student obtained trait scores of two fours and two threes, the student's score would be considered at the cut for identifying ready performance on the spring 2014 base Writing form. This pattern of four trait scores corresponded to a scale score of 428 on the spring 2013 base Writing form. A scale score of 428 is used as the cut score for writing at all grade levels. The writing Benchmark may be updated as more longitudinal data become available.

ACT Readiness Benchmarks for ELA and STEM

The ACT Readiness Benchmarks for ELA and STEM are computed as the average of the subject Readiness Benchmarks that contribute to each score. ACT Readiness Benchmarks for ELA are the average of the benchmarks for English, Reading, and Writing tests. ACT Readiness Benchmarks for STEM are the average of the benchmarks for Mathematics and Science tests. Table 5.4 lists the ACT Readiness Benchmarks for ELA and STEM across grade levels. Benchmarks could have been established through a separate backmapping process, but the average was used because it was anticipated to be simpler for users and less likely to lead to confusion.[9]

Table 5.4. ACT Readiness Benchmarks in ELA and STEM, Grades 3–10

Grade	ELA	STEM
3	419	416
4	421	418
5	422	420
6	423	422
7	424	424
8	425	426
9	426	429
10	428	432

[9] It is possible for a student to be inconsistent with readiness in a particular subject area. This is due to the compensatory nature of the ELA and STEM scores, where subject area scores are combined to obtain ELA and STEM scores. For example, a student could be below the mathematics benchmark and still be above the STEM benchmark if the student had a relatively high Science test score. In such a case, the Science test score could pull the STEM score up enough to meet the STEM benchmark.

ACT Readiness Levels

In addition to the ACT Readiness Benchmarks, students are provided with a description of where they perform relative to the ACT Readiness Benchmarks in English, mathematics, reading, science, or writing. This description is called the ACT Readiness Level and includes four categories defined by three cut scores for each subject and each grade: high cut above the benchmark, the benchmark, and low cut below the benchmark. The high and low cuts were set considering the standard error of measurement (SEM; Geisinger 1991). Specifically, the high cut was set to be two SEMs above the Benchmark, and the low cut was set to be two SEMs below the Benchmark for all subjects except writing.

Two SEMs were chosen because it represented a substantial deviation from *Ready*. From a statistical perspective, under typical assumptions, two standard errors from the benchmark represent a roughly 95% confidence interval for the *Ready* category; we can be about 95% confident that scores falling outside of this range are indeed above or below the *Ready* cut. Of course, values other than two SEMs could have been chosen to define the ACT Readiness Levels, but two SEMs was deemed a reasonable compromise between adequate representation of the descriptors *Exceeding* and *In Need of Support* and statistical characteristics indicative of deviation from *Ready*.

For writing, once the ACT Readiness Benchmark was established, the low and high cuts were set by content experts based on rubrics. Specifically, the low cut was defined as the scale score on the spring 2013 base Writing test form that corresponded to two threes and two twos on the four domains. The high cut was defined as the scale score on the spring 2013 base Writing test form that corresponded to two fives and two fours on the four domains. All benchmarks, low cuts, and high cuts are presented in table 5.5.

Using the ACT Readiness Benchmarks, performance is classified into four ACT Readiness Levels based on student scale scores:

1. Exceeding: at or above the high cut
2. Ready: at or above the benchmark and below the high cut
3. Close: at or above the low cut and below the benchmark
4. In Need of Support: below the low cut

Each of the four ACT Readiness levels is intended to provide a description of where students perform relative to the ACT Readiness Benchmarks in each subject area.

Table 5.5. Benchmarks, Low Cuts, and High Cuts for ACT Readiness Benchmarks by Subject and Grade

Subject	Tested Grade	Low Cut	Benchmark	High Cut
English	3	408	413	418
	4	411	417	423
	5	412	419	426
	6	413	420	427
	7	413	421	429
	8	415	422	429
	9	419	426	433
	10	421	428	435
Mathematics	3	409	413	417
	4	411	416	421
	5	412	418	424
	6	414	420	426
	7	416	422	428
	8	419	425	431
	9	422	428	434
	10	426	432	438
Reading	3	411	415	419
	4	412	417	422
	5	415	420	425
	6	416	421	426
	7	417	423	429
	8	418	424	430
	9	419	425	431
	10	422	428	434
Science	3	414	418	422
	4	415	420	425
	5	417	422	427
	6	418	423	428
	7	420	425	430
	8	422	427	432
	9	424	430	436
	10	426	432	438
Writing	3	420	428	436
	4	420	428	436
	5	420	428	436
	6	420	428	436
	7	420	428	436
	8	420	428	436
	9	420	428	436
	10	420	428	436

ACT Readiness Ranges

ACT Readiness Ranges are reported from grade 3 through EHS in English, mathematics, reading, science, writing, ELA, and STEM. ACT Readiness Ranges are based on an ACT Readiness Benchmark that is provided for each assessment. Students who score at or above this benchmark are on target to meet the ACT College Readiness Benchmark in spring of grade 11. ACT Readiness Ranges include the range of ACT Aspire scale scores between the ACT Readiness Benchmark and the maximum scale score possible for a particular subject at a particular grade level. A student's ACT Aspire scale score can be compared to the ACT Readiness Benchmark and ACT Readiness Range to determine whether the student is on target to be Ready.

Readiness Ranges for Reporting Categories

In order to provide students with more detailed information within each subject, items that measure the same skills and abilities are grouped into reporting categories. For each reporting category, the total number of points possible, the total number of points a student achieved and the percentage of points correct are provided.

In addition, the ACT Readiness Range in each reporting category is provided to show where a student who has met the ACT Readiness Benchmark in a particular subject area would typically perform within the reporting category. In this way students can compare the percentage of points in each category to the percentage of points attained by a typical student who is on track to be *Ready*. If their scores fall below the ACT Readiness Range, they may be in need of additional support.

The minimum of the ACT Readiness Range for each reporting category corresponds to the predicted percent correct on that reporting category at the ACT Readiness Benchmark for that subject test. The maximum corresponds to 100% correct. Regression was used to predict the percent correct on the reporting category using the scale score at the Benchmark point.

CHAPTER 6

ACT Aspire Growth

ACT Aspire supports interpretations of student and group-level aggregate growth. Score reports include the following components, each of which contributes to interpretations of growth:

- A student's current and prior-year scores in all tested subjects (English, mathematics, reading, science, and writing)
- Comparison to **ACT Readiness Benchmarks** (chapter 5) that indicate whether students are on target to meet the ACT College Readiness Benchmarks in grade 11
- Classification of a student into **ACT Readiness Levels** (chapter 5) that describe where a student scores relative to the ACT Readiness Benchmarks
- **Predicted score paths** that provide ranges for a student's expected scores in future years (and predicted score ranges on The ACT for grades 9 and 10)
- Starting in spring 2015, classification of student growth as "Low," "Average," or "High," based on **student growth percentiles** (SGPs).

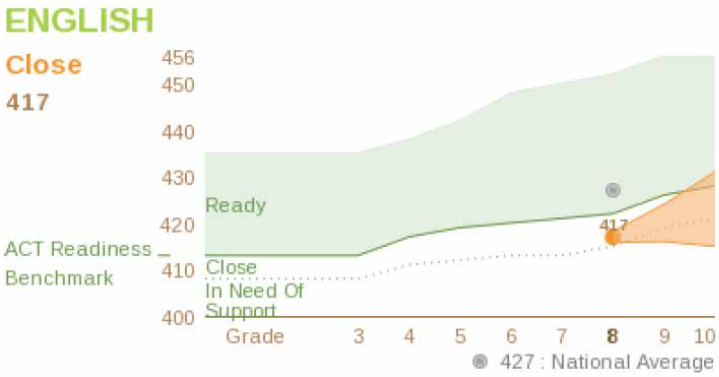
The latter two components (predicted score paths and SGPs) are described in this chapter. Other topics included in this chapter include ACT Aspire gain scores, growth-to-standards models, measurement error of growth scores, and aggregate growth scores for research and evaluation.

Predicted Score Paths

Predicted score paths are reported to enhance understanding of where a student (or a group of students) is likely to score in future years, assuming typical growth. This information can be used:

- To determine if students are likely to meet ACT Readiness Benchmarks over the next two years

Figure 6.1. Prototype of ACT Aspire Student Progress Report



- To identify students who are unlikely to meet a future-year standard (other than the ACT Readiness Benchmark) and thus are candidates for extra academic support
- To predict aggregate future achievement for a classroom, school, district, or state
- To predict ACT score ranges (for grades 9 and 10)

Student Progress Reports

One-year predicted paths are based on the estimated 25th and 75th percentile of the test score distribution, conditional on the prior-year test score. The full predicted path, which encompasses one-year and two-year predictions, is drawn by extending the one-year predictions for another year in a linear fashion. A prototype Student Progress Report that illustrates the predicted path is shown in figure 6.1. The predicted path is represented by a cone-shaped orange-shaded area that covers two years. In this example, an eighth-grade student scored 417 on the ACT Aspire English assessment. The student's predicted path covers the score range 416–424 for grade 9 and 415–431 for grade 10. (Note that the numbers forming the predicted score ranges do not appear on the progress report).

Aggregate Progress Reports

One-year predicted mean scores are used to form aggregate predicted paths for classrooms, schools, and districts. The aggregate predicted paths are drawn as lines connecting the current year mean score to the next year's predicted mean score. Predicted mean scores for classrooms, schools, and districts are calculated as the mean of individual student predicted scores. Individual student predicted scores are based on the estimated 50th percentile of the test score distribution, conditional on the prior-year test score. A prototype Aggregate Progress Report is shown in figure 6.2. In this example, the mean ACT Aspire Science score was 418 for a group of grade 9 students. The predicted grade 10 mean ACT Aspire Science score for the same group is plotted, and the predicted grade 11 mean Science Test score on The ACT is 17.8. We summarize the methods used to develop the predicted paths later in this chapter. The development of the predicted paths is fully documented in a separate report (Allen, forthcoming).

Figure 6.2. Prototype ACT Aspire Aggregate Progress Report

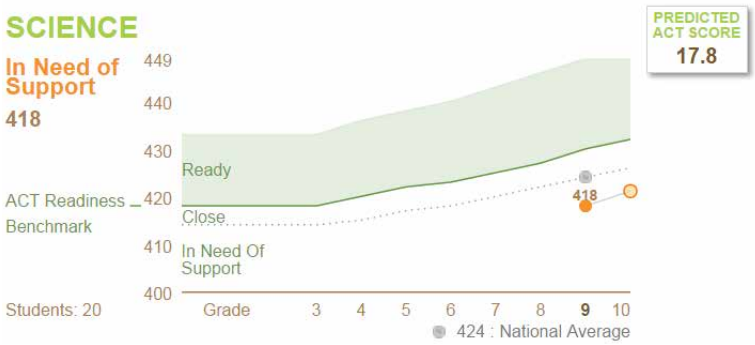


Table 6.1. Longitudinal Samples Used to Develop Predicted Paths in 2014

Grade Level Pair	Subject Area					
	English	Mathematics	Reading	Science	Writing	Composite
3–4	3,784	8,843	8,943	3,403	469	
4–5	3,912	8,724	8,694	3,070	487	
5–6	3,563	7,982	8,008	3,441	431	
6–7	2,329	5,876	5,880	1,677	381	
7–8	1,924	4,583	4,579	2,132	262	
8–9	53,537*	53,537*	53,537*	53,537*	116	53,537*
9–10	172,339*	*172,339*	172,339*	172,339*	0	172,339*
9–11 (The ACT)	50,656*	50,656*	50,656*	50,656*	0	50,656*
10–11 (The ACT)	3,992	3,774	3,540	3,525	922	2,851

*Concordance-derived sample

Samples Used to Develop the Predicted Paths Used for 2014 Reporting

Students who tested in adjacent years form a longitudinal sample of ACT Aspire-tested students and were used to develop predicted paths that are used for ACT Aspire reports. Each year, the longitudinal sample used to develop the predicted paths will be updated with data from the most recent testing year. Here, we describe the sample used in 2014 for the initial development of the predicted paths. For grades 10–11, the longitudinal sample was formed by matching ACT Aspire records from spring 2013 to The ACT records from spring 2014. Because the longitudinal sample for grades 8–9, 9–10, and 9–11 was small, longitudinal samples of students tested with ACT Explore, ACT Plan, and The ACT were used for those grade levels. The ACT Explore and ACT Plan scores were converted to ACT Aspire scores using the EPAS (which includes ACT Plan/ACT Explore) to ACT Aspire concordance tables (see appendix B). Table 6.1 provides the sample sizes for each grade level and subject area. Because of the smaller sample sizes for the Writing tests, data were combined across grade levels for purposes of developing the predicted paths.

Statistical Model Used to Develop the Predicted Paths

Quantile regression (Koenker 2005) was used to estimate the 25th, 50th, and 75th percentiles of the test score distribution, conditional on the prior-year test score. Quantile regression is conceptually similar to ordinary least-squares regression, which is used to estimate the mean of an outcome (denoted Y) given a set of predictor variables (denoted X). Quantile regression estimates selected quantiles of an outcome variable given a set of predictor variables.

Coverage Rates of the Predicted Paths

Using an inclusive definition for predicted path coverage (scores greater than or equal to the lower score of the projected range and scores less than or equal to the upper score of the range), more than 50% of test scores are expected to fall within the one-year predicted path. Less than 25% of test scores should be above the one-year predicted path score range and less than 25% should be below. Coverage rates of the one-year predicted paths were calculated using the longitudinal sample of ACT Aspire and ACT-tested students. For grade level pairs 3–4 through 7–8, the coverage rates ranged from 54% to 57% for English, 55% to 59% for Mathematics, 56% to 57% for Reading and Science, and 55% to 60% for Writing. For grade 10 (ACT Aspire) to 11 (The ACT), coverage rates were 56% for English, 62% for Mathematics, 57% for Reading, 58% for Science, and 73% for Writing. Coverage rates for grade level pairs 8–9, 9–10, and 9–11 will be estimated when more ACT Aspire longitudinal data are available.

Limitations of the Predicted Paths

The predicted paths for grades 3–7 were developed using samples of students tested in spring 2013 and spring 2014, with a large percentage coming from one state (Alabama, the first state to administer ACT Aspire on a statewide basis). With larger samples of students, the estimation of the predicted paths will have less sampling error. It is also possible that predicted path estimates could shift up or down with the inclusion of more school districts and states in the sample. The predicted paths will be re-estimated as more data become available.

The predicted paths for grades 8 and 9 were developed using large samples of students tested with the ACT EPAS system, with greater geographic diversity. However, this approach relies on the EPAS to ACT Aspire concordance, which could introduce bias into the estimation of the predicted paths. As more data become available, the predicted paths for grades 8 and 9 will be estimated using ACT Aspire data without the use of the concordance. The two-year predicted score ranges, defined as a linear extension of the one-year ranges, will likely have asymmetric coverage. Future ACT Aspire reports may include two-year predicted paths that are nonlinear extensions of the one-year paths and may have greater prediction accuracy.

Student Growth Percentiles

Student growth percentiles (SGPs) represent the relative standing of a student's current achievement compared to others with similar prior achievement. ACT Aspire SGPs, ranging from 1 to 100, will be reported starting in spring 2015 for students who also tested in spring 2014. They are used to classify students' growth into the following categories: "Low" (SGP<25), "Average" (SGP between 25 and 75), or "High" (SGP>75). ACT Aspire SGPs measure growth over one-year time intervals.

SGPs will be estimated using quantile regression methods (Koenker 2005) by the SGP R package (Betebenner, Vanlwaarden, Domingue, and Shang 2014). SGPs will be reported annually for students who also tested the previous year with the tests one grade level apart. While the minimum requirement will be the prior-year ACT Aspire score in the same subject area, the model will use up to three years of prior scores when available.

When interpreting SGPs, the reference group used to estimate the model should always be considered. The SGPs used for ACT Aspire will be based on all students tested nationally in adjacent years, so the reference group is expected to change over time. Because of these expected changes and possible changes in the amount of student growth observed nationally over time, SGPs should be interpreted as a measure of relative growth. Because ACT Aspire scale scores are equated across different test forms (see chapter 11), the scale scores maintain similar properties and can be used to make interpretations about growth that are not dependent on a reference group.

ACT Aspire Gain Scores

Each ACT Aspire subject area except writing shares a common scale across grade levels (see chapter 1 for details on the scaling procedures), making it possible to compare scores over time on the same scale. Gain scores are the arithmetic difference in scores from one year to the next. Gain scores are an attractive growth measure because of their simplicity and intuitive appeal. Unlike SGPs, the interpretation of gain scores is independent of reference groups.

For all subjects except writing, positive ACT Aspire gain scores are anticipated because students are expected to increase their knowledge and skills in the tested areas after one year of schooling. For ACT Aspire Writing, gain scores are less meaningful because the writing scale does not share a common scale across grade levels (see chapter 1). For all subjects, including writing, viewing scores graphically over multiple years provides insights about a student's current achievement level, as well as progress made over multiple years with respect to ACT Readiness Benchmarks and ACT Readiness Levels.

Table 6.2. ACT Aspire Gain Score Means and Standard Deviations

Grade Level Pair	Subject Area					
	English	Mathematics	Reading	Science	Writing	Composite
3–4	4.3 (5.3)	4.0 (3.7)	3.4 (4.1)	3.9 (4.6)	−0.2 (6.0)	
4–5	3.5 (5.5)	3.2 (4.4)	3.3 (4.4)	3.1 (4.8)	0.4 (6.5)	
5–6	2.7 (6.0)	3.4 (4.8)	3.5 (4.9)	2.6 (5.0)	5.9 (8.1)	
6–7	2.4 (6.8)	1.6 (5.4)	1.5 (5.0)	2.3 (5.3)	−3.3 (7.0)	
7–8	2.8 (6.6)	2.9 (5.4)	4.3 (5.3)	4.6 (5.4)	−2.8 (6.2)	
8–9	2.3 (5.9)	1.7 (5.1)	1.5 (5.1)	1.7 (5.6)	−1.0 (5.4)	1.8 (3.4)
9–10	2.5 (5.9)	2.7 (5.4)	2.7 (5.3)	2.0 (5.9)		2.5 (3.4)

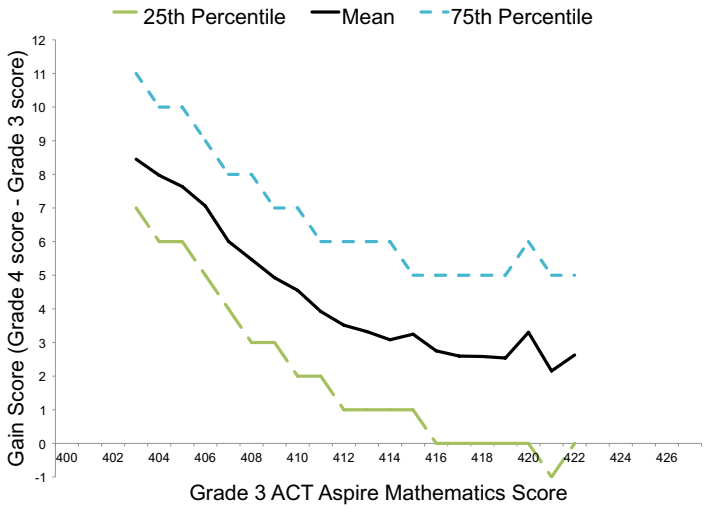
For the longitudinal samples used to develop the ACT Aspire predicted paths (see table 6.1), gain score means and standard deviations are provided in table 6.2. There is considerable variation across grade levels and subject areas in mean gain scores. For all subjects except writing, mean gain scores are always positive, showing that students in the sample typically increased their knowledge and skills in the tested areas after one year of schooling.

For the Writing test, one should not necessarily expect positive mean gain scores because the scale is not the same across grade levels. The mean gain scores for ACT Aspire Writing ranged from 5.9 for grades 5–6 to −3.3 for grades 6–7. Because of the smaller sample sizes for Writing, the mean gain scores reported in table 6.2 are subject to greater sampling error. Writing scores increased substantially from grade 5 to grade 6 (mean gain score of 5.9), and then declined from grade 6 to grade 7 (mean gain score of −3.3) and from grade 7 to grade 8 (mean gain score of −2.8). One possible explanation for the increase in scores for grade 6 is that the grade 6 Writing test is a narrative writing exercise, unlike the grades 4 and 7 tests, which are exercises in expository writing, and the grades 5 and 8 tests, which are exercises in persuasive argumentation. The data suggest that students performed better on the grade 6 narrative writing exercise, perhaps because of greater comfort and familiarity with the narrative mode. At grade 6, the ACT Aspire Writing score range expands from 408–440 to 408–448, with a maximum raw score of 6 (instead of 5) for each domain score. It is likely that this scale increase is also partly responsible for the increase in scores from grade 5 to grade 6.

Interpreting ACT Aspire Gain Scores

Because ACT Aspire scores in all subject areas except writing are reported on a common scale, gain scores are intended to be interpreted as measuring change in knowledge and skills from one year's test to another. However, because no educational scale can be interpreted as having equal intervals in a strict sense, it should not be assumed that the meaning of gain scores is the same across the score

Figure 6.3. ACT Aspire gain score statistics for Grade 3–4 Mathematics



scale. For example, it is possible that some regions of the score scale can be more sensitive to student learning. Also, students who score at the low end of a scale are generally expected to have larger one-year gain scores than students who score at the high end. This phenomenon is illustrated in figure 6.3 using the longitudinal sample of ACT Aspire-tested students. Students in this sample took the Grade 3 Mathematics test in spring 2013 and the Grade 4 Mathematics test in spring 2014. The figure shows the mean gain score, as well as the 25th and 75th percentile of the gain score distribution, for each Grade 3 Mathematics score point with a sample size of at least 50 students. The negative relationship between prior-year score and one-year gain score results because test scores are only estimates of true achievement levels: students who score far above (below) the mean are more likely than others to have scored above (below) their true achievement level and tend to score closer to the mean the next year.

Growth-to-Standards Models with ACT Aspire

Unlike normative growth measures such as SGPs, growth-to-standards models determine if students are making sufficient progress toward a performance standard, such as the ACT Readiness Benchmarks or other ACT Readiness Levels. ACT Aspire supports growth-to-standards models because scores are reported over time and plotted against the ACT Readiness Benchmarks and ACT Readiness Levels (see figure 6.1). The predicted paths also support growth-to-standards models because they indicate how students are likely to perform over the next two years, assuming typical growth. For example, if a fourth grader's predicted path falls below the ACT Readiness Benchmark for grades 5 and 6, he or she knows that atypically high growth over the next two years will be needed to reach the performance standard.

Growth-to-standards models can be implemented by specifying a performance standard such as the ACT Readiness Benchmark and specifying the amount of time

students have to reach the performance standard. For the student represented by the example report in figure 6.1, the student's score (417) was 5 points below the grade 8 ACT Readiness Benchmark for English (422). If the model assumes students have two years to catch up, the student would need to gain 11 score points to reach the grade 10 ACT Readiness Benchmark for English (428). Under a growth-to-standards model, the student would need to gain at least 6 points from grade 8 to grade 9 to have made sufficient progress toward the grade 10 performance standard. SGPs can also be used within a growth-to-standards model to communicate how much progress is needed using the growth-percentile metric instead of the gain-score metric.

Measurement Error of Growth Scores

Measures of individual student growth, including SGPs, are subject to measurement error. This means that a student's actual growth in academic achievement may be different than what is represented by their SGP (Wells, Sireci, and Bahry 2014) or gain score. All test scores have measurement error, and measurement errors of SGPs and gain scores are more pronounced because multiple test scores are involved. In tests with vertical scales such as ACT Aspire, negative gain scores may be observed, due in part to measurement error. Because of their measurement error, neither SGPs nor gain scores should be used as the sole indicator of a student's academic progress over one year.

Aggregate Growth Scores for Research and Evaluation

For classrooms, schools, districts, states, and other user-defined groups, aggregate growth statistics are available to describe how much growth occurred in each group. Mean scale scores are plotted on longitudinal progress reports and the percentage of students in each growth category (low, average, or high) can be reported. It is also possible to calculate other summary growth measures, such as the median SGP, using available data.

Data from ACT Aspire may be used as one indicator of program effectiveness. One of the secondary interpretations of ACT Aspire scores involves providing empirical data for inferences related to accountability (see chapter 9 or ACT 2014b). Before using ACT Aspire for evaluating program effectiveness, a content review is critical to ensure that ACT Aspire measures important and/or relevant outcomes of the program. For example, a district might want to use the multiyear change in median SGP of 10th grade English students as one measure of the effectiveness of a new 10th grade English curriculum. Prior to implementation, ACT Aspire content should be reviewed against the intended outcomes of the program to evaluate its appropriateness for measuring curriculum effectiveness. If used for accountability, ACT Aspire growth data should be one of multiple sources of evidence regarding student performance for particular uses. As the stakes for particular uses increase, it becomes more important to carefully evaluate ACT Aspire score interpretations for these uses and to gather multiple sources of evidence.

CHAPTER 7

Progress toward Career Readiness

Introduction

For decades it has been a commonly held belief that high school students planning to go to college need to take more rigorous coursework than those going directly into the workforce. Today, however, many employers report that in an expanding global economy, entry-level workers need many of the same types of knowledge and skills as college-bound students (ACT 2006). To help students predict and monitor whether they are on track to be career ready toward the end of high school using ACT Aspire performance, students at grade 8 through EHS with Composite scores (i.e., receiving scores on English, Mathematics, Reading, and Science tests) receive a Progress toward Career Readiness indicator.

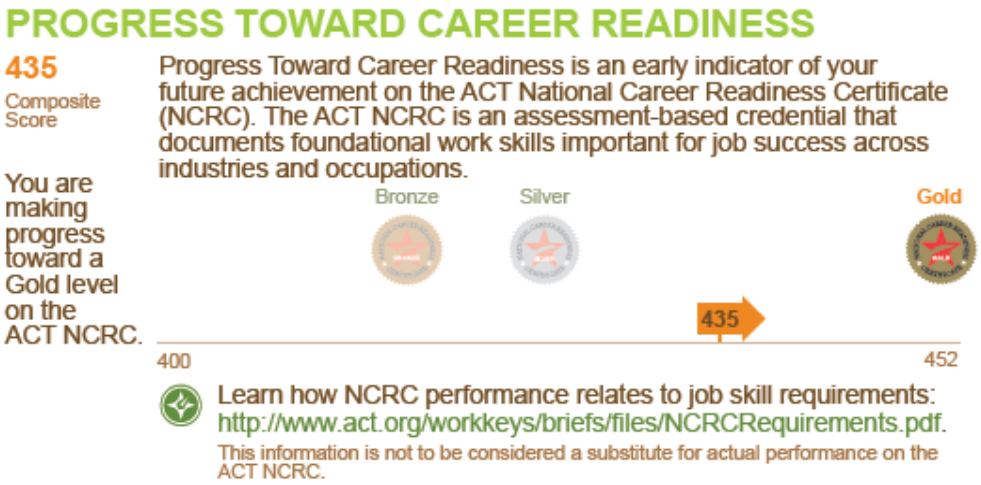
The ACT National Career Readiness Certificate™ (ACT NCRC®), awarded based on ACT WorkKeys® test results, is a portable credential that demonstrates achievement and is based on assessment level scores associated with workplace employability skills in three areas: applied mathematics, locating information, and reading for information. Assessment level scores typically run from level 3 to level 6 or 7. In addition, the ACT NCRC has four levels of certificate: Bronze, Silver, Gold, and Platinum. A Platinum certificate is earned if the three assessment level scores a student earns are all level 6 or higher. A Gold certificate is earned if the three assessment level scores a student earns are all level 5 or higher. A Silver certificate is

earned if the three assessment level scores a student earns are all level 4 or higher. A Bronze certificate is earned if the three assessment level scores a student earns are all level 3 or higher. If any of the three assessment level scores are less than level 3, no certificate is achieved (ACT 2007a).

The actual knowledge and skills required across careers and career domains may differ greatly. Further, The ACT and ACT Aspire do not measure all such knowledge and skills. Rather, The ACT and ACT Aspire measure academic achievement, or foundation skills which are a subset of skills associated with career readiness. The linkage between ACT Aspire and the ACT NCRC is based only on these academic skills and provides a prediction of future performance. Because the constructs and content across The ACT and ACT Aspire differ somewhat from that of the ACT WorkKeys used in the ACT NCRC, a prediction method was used to indicate if a student is likely to meet Bronze, Silver, or Gold ACT NCRC level at the completion of high school.

The Progress toward Career Readiness cuts were created for grades 8–11 to be used as indicators of progress towards career readiness. These cuts were used as an indicator of predicted future ACT NCRC performance (see figure 7.1). The first step in creating the cuts was linking ACT NCRC level to the EPAS scale. Prediction was used to link the EPAS Composite score scale to the ACT NCRC levels. A concordance was then used to find the ACT Aspire scores that corresponded to each of the cut scores on the EPAS scale. Finally, backmapping was used to obtain the cut scores associated with Bronze, Silver, and Gold ACT NCRC levels for grades 8–10.

Figure 7.1. Sample Progress toward Career Readiness indicator from ACT Aspire report



Link ACT NCRC Levels to the EPAS Composite Scores

Data from over 110,000 grade 11 students who took ACT WorkKeys (and obtained an ACT NCRC) and The ACT were used to establish the link between the EPAS Composite and the ACT NCRC. Four separate logistic regressions were conducted to determine the ACT Composite scores that corresponded to the 50% chance of obtaining each ACT NCRC level. The ACT Composite scale score was the independent variable, and the status of whether a student was at or above a specific ACT NCRC level was the dependent variable (similar methods have been used for setting ACT College Readiness Benchmarks; e.g., Allen and Sconing 2005).

Table 7.1 provides descriptive statistics of the Composite scores on The ACT for students at different ACT NCRC levels. As ACT NCRC level increases, the mean and median of ACT Composite scores increase, which is indicative of the positive relationship between ACT NCRC level and ACT Composite scores. The variability in performance on The ACT increases as ACT NCRC level increases through the Gold level. Between the Gold and Platinum levels, variability decreases slightly, which is likely the result of instability in the standard deviation due to the relatively small number of students in the Platinum group compared to the others.

Table 7.2 presents the logistic regression results—unrounded and rounded Composite scores on The ACT that are associated with the 50% chance of obtaining different ACT NCRC levels. For example, an ACT Composite scale score of 17 is required for the 50% chance of obtaining the ACT NCRC Silver certificate or higher while an ACT Composite scale score of 25 is required for a 50% chance of obtaining the ACT NCRC Gold certificate or higher.

Table 7.1. Descriptive Statistics of ACT Composite Scale Scores by Each ACT NCRC Level

ACT NCRC levels	ACT Composite Scores					
	N	Mean	Median	SD	Min	Max
No Certificate	13,628	13.69	13	2.28	4	32
Bronze	23,726	15.86	16	2.64	6	33
Silver	50,087	19.76	20	3.56	7	35
Gold	25,669	25.20	25	3.81	11	36
Platinum	1,497	30.33	31	3.15	18	36

Table 7.2. ACT Composite Scale Scores Indicating a 50% Chance of Obtaining Different ACT NCRC Certificates

ACT NCRC levels	ACT Composite Cut Scores	
	Unrounded	Rounded
Bronze	12.74	13
Silver	16.43	17
Gold	24.15	25
Platinum	34.80	35

Table 7.3. Descriptive Statistics of the Analysis Sample (*N*=13,528)

	EPAS Scale Score	ACT Aspire Scale Score
Mean	16.66	424.05
SD	3.95	7.72
Correlation	.86	

To ensure each score on The ACT corresponds to at least a 50% chance of receiving an ACT NCRC level, all unrounded Composite scores were rounded up to the final reported cut values. These four cut values are based on a sample of grade 11 students, so they should be interpreted as an indicator of career readiness for grade 11 students.

Link the EPAS Composite Scores to ACT Aspire Composite Scores

The second step in the linkage between ACT Aspire and ACT NCRC certificate levels was to establish a concordance between the EPAS Composite scores on the EPAS tests (ACT Explore, ACT Plan, or The ACT) and ACT Aspire Composite scores. Data from spring 2013 ACT Aspire tests were merged with historical EPAS student data files to obtain a sample of students with EPAS and ACT Aspire Composite scores. Table 7.3 presents means, standard deviations, and correlations for the 13,528 students included in the sample.

The concordance relationship between EPAS and ACT Aspire scales was estimated using the equipercentile method described by Kolen and Brennan (2014). The same methodology that was used to find the concordance relationship between EPAS and ACT Aspire for the individual subjects was used to find the relationship between EPAS and ACT Aspire Composite scores. See chapter 4 for details on the methodology.

Identify the ACT Aspire Composite Scores Corresponding to Each ACT NCRC Level

Once the concordance between the EPAS Composite scores and the ACT Aspire Composite scores was established, the ACT Aspire scale score that best predicts each ACT NCRC level for grade 11 was taken directly from the concordance table. Table 7.4 presents the EPAS Composite scale scores corresponding to ACT NCRC Levels and their concorded ACT Aspire Composite scale scores. The Platinum level was excluded due to small sample sizes.

To obtain the Progress toward Career Readiness indicator for each ACT NCRC level for grades 8, 9, and 10, a backmapping procedure was adopted using *z*-scores and the grade 11 Progress toward Career Readiness indicator as the starting point. (The same procedure was used to backmap the ACT Readiness Benchmarks; see chapter 5.) The ACT Aspire Composite score associated with each ACT NCRC level for a particular grade was the scale score that corresponds to the same standardized score (i.e., *z*-score) as the *z*-score of the EPAS Composite for each grade 11 ACT NCRC level in the sample used for the EPAS to ACT Aspire concordance analysis. Grade 11 *z*-scores associated with Bronze, Silver, and Gold ACT NCRC levels were matched to the *z*-scores at grades 8, 9, and 10. Then the ACT Aspire scale score associated with each *z*-score was identified to determine the backmapped Progress toward Career Readiness indicator corresponding to Bronze, Silver, and Gold. Table 7.5 presents the ACT Aspire scale score for each of the three ACT NCRC levels for grades 8–10.

Table 7.4. EPAS and ACT Aspire Composite Scores Corresponding to ACT NCRC Levels

ACT NCRC Level	Corresponding EPAS Composite	Corresponding ACT Aspire Composite (Grade 11)
Bronze	13	416
Silver	17	425
Gold	25	439

Table 7.5. ACT Aspire Composite Scores Corresponding to the 50% Chance of Obtaining Each ACT NCRC Level

	ACT NCRC Level		
Grade	**Bronze**	**Silver**	**Gold**
8	415	422	434
9	415	423	436
10	416	425	439

Report Progress toward Career Readiness using ACT Aspire Cut Scores

Students who take the four 8th-grade or EHS tests English, Mathematics, Reading, and Science obtain a Composite score and receive an indicator of their Progress toward Career Readiness (see figure 7.1 for a sample from the score report). The student's Composite score is listed along with the Bronze, Silver, and Gold level ACT NCRC certificates. The indicator reported to the student is a verbal statement indicating which certificate level the student is making progress toward. A purpose of the indicator is to encourage students to think about the knowledge and skills future job training will require. Whether or not a student plans to enter college immediately after high school, there is likely a need for some type of postsecondary training. It is not possible to profile the requirements for all possible postsecondary paths a student may consider, but the ACT NCRC and the ACT WorkKeys job profiles provide a starting point for students to understand that even if they are not planning to go to college, there are still knowledge and skills they need to attain before they leave high school. The specific foundational knowledge and skills required across careers and career paths differ, and ACT Aspire may not measure all such skills.

The Progress toward Career Readiness indicator provides a statistical prediction of the likely ACT NCRC level a student would obtain toward the end of high school given the student's current performance, but caution should be used in its interpretation. The Progress toward Career Readiness indicator is *not* a substitute for actual ACT NCRC level obtained by taking ACT WorkKeys. Actual performance could differ from the statistically predicted performance for a variety of reasons, including such factors as statistical uncertainty in the prediction, a student's individual educational achievement, and a student's growth trajectory.

CHAPTER 8

ACT Aspire Reliability

Some degree of inconsistency or error is contained in the measurement of cognitive characteristics. A student administered one form of a test on one occasion and a second, parallel form on another occasion likely would earn somewhat different scores on the two administrations. These differences might be due to the student or the testing situation, such as different motivation, different levels of distractions across occasions, or student growth between testing events. Differences across testing occasions might also be due to the particular sample of test items or prompts included on each test form. While procedures are in place to reduce differences across testing occasions, differences cannot be eliminated.

Reliability coefficients are estimates of the consistency, or precision, of test scores. They typically range from zero to one, with values near one indicating greater consistency and those near zero indicating little or no consistency. The standard error of measurement (SEM) is closely related to test reliability. The SEM summarizes the amount of error or inconsistency in scores on a test. The *Standards for Educational and Psychological Testing* states: "For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported" (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] 2014, 43).

Raw Score Reliability

Reliability coefficients are usually estimated based on a single test administration by calculating the inter-item covariances. These coefficients are referred to as internal consistency reliability. Cronbach's coefficient alpha (Cronbach 1951) is one of the most widely used estimates of test reliability and was computed for all of the ACT Aspire tests. Coefficient alpha can be computed using the following formula:

$$\hat{\alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_x^2}\right) = \left(\frac{k}{k-1}\right)\left(\frac{\sum_{i=1}^{k}\sum_{j\neq i}^{k} s_{ij}}{s_x^2}\right)$$

where k is the number of test items, s_i^2 is the sample variance of the ith item, s_{ij} is the sample covariance between item i and item j, and s_x^2 is the sample variance of the observed total raw score.

Although coefficient alpha is often used to estimate internal consistency reliability, it can, at times, underestimate the true value of test reliability depending on the characteristics of the specific test under consideration. In computing test reliabilities for the ACT Aspire tests, stratified coefficient alpha (Cronbach, Schonemann, and McKie 1965) and congeneric reliability (Gilmer and Feldt 1983) coefficients were also computed as a check on coefficient alpha. All three reliability coefficients produced nearly equal values for ACT Aspire subject tests at each grade level. As a consequence, Cronbach's coefficient alpha was used in reporting raw score reliability in tables 8.1 and 8.2.

Table 8.1. Raw Score and Scale Score Reliability Coefficient Ranges by Grade for Four ACT Aspire Subject Tests and the Composite Score: Spring 2013 Special Studies Data

		Grade						
Subject	**Score**	**3**	**4**	**5**	**6**	**7**	**8**	**EHS**
English	**Raw**	.67–.79	.75–.80	.76–.79	.78–.81	.78–.85	.84–.87	.90–.91
	Scale	.70–.79	.76–.80	.77–.79	.80–.81	.79–.85	.84–.86	.90–.91
Math	**Raw**	.73–.79	.55–.76	.57–.77	.61–.74	.66–.83	.84–.87	.86–.89
	Scale	.75–.79	.62–.75	.65–.77	.74–.77	.70–.85	.86–.89	.87–.90
Reading	**Raw**	.83–.85	.83–.84	.81–.84	.81–.84	.77–.83	.81–.86	.87–.87
	Scale	.83–.85	.83–.85	.81–.84	.82–.85	.79–.84	.82–.87	.88–.88
Science	**Raw**	.85–.88	.83–.84	.84–.87	.85–.88	.85–.89	.84–.88	.87–.90
	Scale	.86–.88	.83–.85	.84–.88	.86–.89	.85–.90	.85–.89	.86–.89
Composite*	**Scale**						.95–.96	.96–.97

* Composite scores are not reported below grade 8.

Table 8.2. Raw Score Reliability Coefficient Ranges by Grade for Four ACT Aspire Tests: Spring 2014 Operational Data

	Grade						
Subject	**3**	**4**	**5**	**6**	**7**	**8**	**EHS**
English	.78–.82	.74–.78	.75–.78	.80–.82	.78–.83	.84–.85	.88–.90
Math	.78–.79	.67–.68	.67–.71	.77–.79	.81–.84	.86–.87	.82–.88
Reading	.83–.85	.82–.84	.82–.84	.82–.84	.79–.80	.81–.83	.85–.87
Science	.86–.89	.82–.84	.83–.84	.86–.87	.87–.89	.86–.87	.85–.89

The raw score reliabilities from spring 2013 special studies listed in table 8.1 for English, Mathematics, Reading, and Science tests ranged from .55 (Grade 4 Mathematics) to a high of .91 (EHS English). Mathematics reliabilities tended to be quite low in some cases, particularly grades 4–7.

The raw score reliabilities from the spring 2014 operational administration are listed in table 8.2 for English, Mathematics, Reading, and Science tests. Reliabilities improved compared to 2013 and ranged from .67 (Grades 4 and 5 Mathematics) to .90 (EHS English). Mathematics reliabilities still tended to be relatively low, specifically in grades 4 and 5. Additional analysis reported elsewhere will examine the Mathematics test, especially grades 4 and 5, in more detail.

Tests that consist of a single item typically use different types of information to estimate raw score reliability. For a test like ACT Aspire Writing, which consists of a single writing prompt scored by a single rater, one aspect of score reliability is rater consistency. While rater consistency should not be confused with score reliability due to task (in this case, the particular writing prompt administered), rater consistency is an important contributor to the reliability of writing prompt scores. One way to analyze rater consistency is by estimating correlations between two raters. In the spring 2013 ACT Aspire special studies, a subsample of students at each grade were rated by two raters. Table 8.3 contains Pearson correlations between raters. The writing prompt is rated on four dimensions, or traits, and these traits are combined to obtain a total score for the writing prompt.[10] Correlations are reported for trait scores and total scores between raters. Writing total score correlations ranged from .70 (Grade 7) to .81 (EHS online and Grade 3 online). Trait correlations ranged from .61 to .77.

[10] The four traits include generating ideas, development, organization, and language use. For additional details, see (ACT 2014a or ACT 2014b).

Table 8.3. Writing Test Correlations between Rater 1 and Rater 2, by Trait* and by Form: Spring 2013 Special Studies Data

Grade	N	Trait 1	Trait 2	Trait 3	Trait 4	Average Correlation among the Four Traits	Total Score Correlation
3	2,141	.68	.72	.73	.66	.70	.81
4	1,902	.74	.72	.70	.71	.72	.78
5	2,041	.63	.65	.65	.69	.66	.73
6	2,190	.65	.66	.68	.65	.66	.72
7	2,243	.61	.62	.62	.65	.62	.70
8	1,701	.70	.72	.72	.65	.70	.77
EHS	2,033	.76	.76	.77	.72	.75	.81

* Trait 1 = generating ideas, trait 2 = development, trait 3 = organization, and trait 4 = language use. For additional details, see *ACT Aspire Technical Bulletin #1* (ACT 2014b).

Table 8.4. Writing Test Reliability Coefficients Based on Four Trait Scores: Spring 2013 Special Studies Data

Grade	N	Reliability
3	5307	.91
4	4709	.96
5	5046	.95
6	5407	.96
7	5563	.95
8	4717	.96
EHS	5073	.96

Although ACT Aspire Writing consists of a single writing prompt, ratings on the four traits that contribute to a final writing score can be used to obtain an internal consistency reliability estimate, Cronbach's coefficient alpha, for a writing prompt. These reliabilities are listed in table 8.4 for writing prompts from spring 2013 and ranged from .91 to .96, which indicates that the four traits within a single writing prompt are quite reliable. However, these coefficients are limited because they do not account for rater, prompt, or occasion variability, each of which is likely to be a more important contributor to the precision of the ACT Aspire Writing scores than traits within a prompt. The large internal consistency reliability based on trait scores is a reflection of the large correlations among traits within a writing prompt.

Scale Score Reliability and Conditional Standard Error of Measurement

The ACT Aspire scale was developed using item response theory (IRT), which includes statistical models that can be used to obtain estimates of scale score reliabilities and conditional standard errors of measurement (CSEM). For ACT Aspire, these reliabilities and CSEMs are based on the three-digit reported scale scores.

Under IRT, a parameter for a student's proficiency is commonly represented as θ (theta). In addition to this student parameter, IRT models used for ACT Aspire contain item parameters that represent particular characteristics of the items under the model. Additional details regarding IRT models applied to ACT Aspire can be found elsewhere in ACT Aspire technical documentation, specifically descriptions of the scaling study, or for more information, see Baker and Kim (2004), de Ayala (2009), or Yen and Fitzpatrick (2006).

Using item parameter estimates and an estimated person proficiency distribution ($\varphi(\theta_i)$, where θ_i is the quadrature point), a version of the Lord-Wingersky recursive algorithm (Hanson 1994; Kolen and Brennan 2014, 199) was used to estimate the conditional distribution of the expected raw scores, *x,* given a theta value, $f(X = x \mid \theta_i)$. With a maximum number of raw score points, *K,* and raw-to-scale score conversion, *sc*(*x*), the true scale score given θ_i is

$$\xi(\theta_i) = \sum_{x=0}^{K} sc(x) f(X = x \mid \theta_i),$$

and the conditional standard error of measurement (*CSEM) given* θ_i is

$$CSEM[sc(x \mid \theta_i)] = \sqrt{\sum_{x=0}^{K} [sc(x) - \xi(\theta_i)]^2 f(X = x \mid \theta_i)}.$$

The average error variance of scale scores was calculated as follows,

$$\sigma_{E_s}^2 = \sum_i CSEM[sc(x \mid \theta_i)]^2 \times \varphi(\theta_i),$$

and the true scale score variance ($\sigma_{T_s}^2$) was calculated based on $\xi(\theta_i)$ and $\varphi(\theta_i)$. Finally, the reliability of scale score was estimated as

$$\rho_{IRT_S} = 1 - \frac{\sigma_{E_s}^2}{\sigma_{E_s}^2 + \sigma_{T_s}^2}.$$

For additional details regarding these procedures for estimating CSEM and reliability, see Kolen and Brennan (2014).

The CSEMs of scale scores are listed in figure 1.4 under the second figure for each subject area for ACT Aspire tests for forms administered during spring 2013. Each curve represents a grade level, and each figure contains an ACT Aspire subject. The ACT Aspire scale in each subject was developed to have approximately constant standard errors of measurement throughout the score scale (see chapter 1 on scaling), which implies that scale scores have similar precision for all students. The ACT Aspire score scales begin at 400 and have different maximum value (up to 460) depending on the grade and subject area. In figure 1.4 we see that for most of the ACT Aspire score scale, each of the curves representing the CSEMs for a grade are within a range of two scale score points. The CSEMs drop dramatically at very low scale scores but never appear higher than roughly 4 scale score points. While these CSEMS are not perfectly flat, which would imply a perfectly constant CSEM across the score scale, from a practical perspective they are reasonably consistent. For example, the second graph in figure 1.4 contains CSEMs for English, and the CSEMs for most of the score scales are between 2 and 4 scale score points across all grades. Within each grade, CSEMS are generally within one scale score point for most of the score scale. For example, the grade 4 English CSEM fluctuates between 2 and 3 across most of the score scale (excluding the bottom few scale score points).

The scale score reliabilities are listed in table 8.1 below the raw score reliabilities. English scale score reliabilities ranged from .70 (grade 3) to .91 (EHS). Mathematics scale score reliabilities ranged from .62 (grade 4) to .90 (grade 7). Reading scale score reliabilities ranged from .79 (grade 7) to .88 (EHS). Science scale score reliabilities ranged from .83 (grade 4) to .90 (grade 7). Scale score reliabilities are useful because they are an estimate of the precision of the scores reported to students and used for interpreting test performance. Raw score reliabilities are also useful for obtaining an estimate of score precision, but raw number-of-point scores are not used for interpreting student performance on ACT Aspire. Therefore, where possible, scale score reliabilities are preferable.

Composite scale score reliabilities are also reported in table 8.1 for those grades where Composite scores are reported to students. The Composite scale score is computed as the average of the four test scale scores and is denoted by *Z*. The formula for computing the score composite reliability is

$$rel_{cmpst} = 1 - \frac{\sum_{i=1}^{4} s_i^2 (1 - \hat{\alpha}_i)}{16 s_z^2}.$$

Composite scale score reliabilities ranged from .95 to .97.

The scale score reliabilities were typically quite similar to the raw score reliabilities calculated using Cronbach's alpha. The largest differences across raw and scale score reliabilities were for Mathematics tests in grades 4–6.

CHAPTER 9

ACT Aspire Validity

According to the *Standards for Educational and Psychological Testing*, "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, and NCME 2014, 11). Validation is the process of justifying a particular interpretation or use and may involve logical, empirical, or theoretical components. As stated by Kane (2006):

> Measurement uses limited samples of observations to draw general and abstract conclusions about persons or other units (e.g., classes, schools). To validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the claims being made, and this in turn requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses. Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified (2006, 17).

The potential interpretations and uses of ACT Aspire scores are numerous and diverse. Some interpretations and uses are anticipated and others are not, but each needs to be justified by a validity argument. The purpose of this chapter is to identify the intended uses of ACT Aspire scores and to provide empirical evidence to validate ACT Aspire score interpretations.

ACT Aspire scores include two primary interpretations and three secondary interpretations. The two primary interpretations are to identify students' readiness on (a) a college readiness trajectory and (b) a career-readiness trajectory. The three secondary interpretations are to provide instructionally actionable information to educators, empirical data for inferences related to accountability, and empirical support for inferences about international comparisons. Each of these interpretations is described in additional detail in other technical documentation (ACT 2014b).

Fundamental to these five interpretations is the assumption that ACT Aspire scores are indicative of performance on a particular set of traits in the subject areas assessed by ACT Aspire: English, mathematics, reading, science, and writing. Scores on each subject area test are intended to provide inferences about students' knowledge and skills (achievement) in these subjects. Scores obtained (or some combination of them) are then interpreted as indicators of readiness for college and career and are intended to be used to identify student status on the path to readiness for college and career. Therefore, one aspect of validation for ACT Aspire is gathering evidence that *ACT Aspire scores are indicative of performance in English, math, reading, science, and writing*.

We can draw on several sources of evidence to validate the argument that ACT Aspire scores are indicative of performance in different subject areas. For example, the standards to which the ACT Aspire is built, the test development process, and content descriptions for each subject area provide evidence that the ACT Aspire tests cover the traits intended. Content evidence to support ACT Aspire score interpretations is described in *Technical Bulletin #1* (ACT 2014b). Other sections of this document describe scaling, equating, and scoring, which provides evidence that the process of ACT Aspire scoring supports desired interpretations of scores. In addition, evidence can be gathered from empirical comparisons of ACT Aspire scores and other assessments testing similar traits. Traditionally, this type of evidence is referred to as convergent validity evidence (or evidence based on relations to other variables). The remainder of this chapter will describe two studies that examine the relationships between ACT Aspire scores and scores from other tests intended to measure similar traits.

These two studies are only a small component of the validation process for ACT Aspire. Validation of ACT Aspire score interpretations and uses will continue to be accumulated and draw on multiple sources of evidence, including the other sections of this document, other technical documentation, and additional research studies.

Study 1: Comparison of ACT Explore and ACT Plan Scores to ACT Aspire Scores

In this study the relationships between ACT Explore, ACT Plan, and ACT Aspire scale scores in English, mathematics, reading, and science were compared using samples of students taking ACT Explore and ACT Aspire (grades 8 and 9) or ACT Plan and ACT Aspire assessments (grade 10). Samples included students participating in special 2013 ACT Aspire studies described in chapter 4 who had also taken ACT Plan or ACT Explore separately. The sample sizes and descriptive statistics by grade and subject in each sample are listed in table 9.1. The samples included students from a total of 122 districts and 263 schools.

Table 9.1. Descriptive Statistics for ACT Explore, ACT Plan and ACT Aspire Scale Scores

Subject	Grade	N	Assessment	Mean	SD	Minimum	Maximum
English	8	7,574	ACT Explore	14.62	4.02	1	25
			ACT Aspire	426.17	8.72	400	452
	9	1,563	ACT Explore	15.73	4.38	2	25
			ACT Aspire	426.87	10.75	400	454
	10	4,643	ACT Plan	17.37	4.65	1	32
			ACT Aspire	429.58	11.19	400	456
Mathematics	8	7,803	ACT Explore	14.88	3.29	1	25
			ACT Aspire	421.22	7.54	401	449
	9	1,641	ACT Explore	16.56	3.76	4	25
			ACT Aspire	423.85	8.47	403	451
	10	4,545	ACT Plan	17.82	4.70	1	32
			ACT Aspire	425.10	9.14	403	455
Reading	8	7,594	ACT Explore	14.60	3.82	1	25
			ACT Aspire	420.15	7.29	401	440
	9	1,484	ACT Explore	15.30	4.20	1	25
			ACT Aspire	420.18	7.74	403	442
	10	4,249	ACT Plan	17.10	4.54	1	30
			ACT Aspire	420.95	8.03	403	442
Science	8	7,779	ACT Explore	16.63	3.21	3	25
			ACT Aspire	421.82	7.75	402	446
	9	1,575	ACT Explore	17.69	3.54	5	25
			ACT Aspire	423.90	8.17	402	446
	10	4,236	ACT Plan	18.59	3.89	1	32
			ACT Aspire	423.97	8.80	401	449
Composite	8	6,419	ACT Explore	15.35	3.16	3	25
			ACT Aspire	422.62	6.95	406	443
	9	1,181	ACT Explore	16.88	3.44	8	25
			ACT Aspire	424.75	7.80	407	445
	10	3,154	ACT Plan	17.89	3.97	6	31
			ACT Aspire	425.22	8.38	408	447

Table 9.2. Correlations between ACT Explore/ACT Plan Scale Scores and ACT Aspire Scale Scores

Sample/Grade	English	Mathematics	Reading	Science	Composite
ACT Explore-ACT Aspire/8	.75	.72	.70	.69	.85
ACT Explore-ACT Aspire/9	.76	.75	.66	.70	.82
ACT Plan-ACT Aspire/10	.78	.77	.65	.69	.84

ACT Explore and ACT Plan are both intended to measure academic achievement, similar to ACT Aspire. In fact, ACT Aspire is intended to serve similar purposes and share similar interpretations as ACT Explore and ACT Plan: ACT Aspire is the successor to those assessments. ACT Explore and ACT Plan scale scores are intended to be on the same scale but have different scale ranges. ACT Explore scale scores range from 1 to 25 and ACT Plan scale scores range from 1 to 32. ACT Aspire scores are also intended to be on the same scale and have different scale score ranges across grades (see table 1.13).

Table 9.1, lists the mean, standard deviation, minimum, and maximum of scale scores observed for ACT Explore/ACT Plan and ACT Aspire. The mean scale scores increase for ACT Explore/Plan and ACT Aspire across grades 8–10. This pattern is consistent with the argument that students at higher grades have higher achievement and therefore should have higher scores.

Table 9.2 lists the correlations between same-subject ACT Explore or ACT Plan scale scores and ACT Aspire scale scores for grades 8–10 (sample sizes are listed in table 9.1). Correlations ranged from .65 (Grade 10 Reading) to .85 (Grade 8 Composite), which are indicative of moderate to strong linear relationships between ACT Explore/ ACT Plan and ACT Aspire. From a linear regression perspective, we can say that 42% to 72% of the variance in scale scores is shared between ACT Explore/ACT Plan and ACT Aspire.

Table 9.3 lists the disattenuated correlations between ACT Explore or ACT Plan and ACT Aspire (sample sizes are listed in table 9.1). Disattenuated correlations are estimates of the linear relationships between scores after taking into account the reliability of each test.[11] In classical test theory, disattenuated correlations are referred to as estimates of the relationship between *true scores*; they provide an estimate of the relationship between ACT Explore/ACT Plan and ACT Aspire as if each contributing score were perfectly reliable. Published or available reliability coefficients for ACT Explore (ACT 2013a), ACT Plan (ACT 2013b) and ACT Aspire (see table 9.4) were used to calculate disattenuated correlations, which ranged from .76 to .92 across subjects, indicative of moderate to strong correlations. From a linear regression perspective, we can say that 58% to 85% of the variance in true scores is shared between ACT Explore/ACT Plan and ACT Aspire.

[11] Disattenuated correlations are calculated as the correlation divided by the square root of the product of the reliabilities or $\rho_{T_X T_Y} = \rho_{XY} / \sqrt{\rho_{XX'} \rho_{YY'}}$.

Table 9.3. Disattenuated Correlations between ACT Explore/ACT Plan Scale Scores and ACT Aspire Scale Scores

Sample/Grade	English	Mathematics	Reading	Science	Composite
ACT Explore-ACT Aspire/8	.89	.88	.82	.83	.89
ACT Explore-ACT Aspire/9	.86	.89	.76	.82	.85
ACT Plan-ACT Aspire/10	.88	.91	.75	.81	.88

Table 9.4. Scale Score Reliabilities

Test and Grade Level	English	Mathematics	Reading	Science	Composite
ACT Aspire 8*	.85	.88	.85	.87	.96
ACT Aspire Early High School*	.91	.89	.88	.88	.97
ACT Explore 8†	.84	.76	.86	.79	.94
ACT Explore 9†	.86	.80	.86	.82	.95
ACT Plan 10‡	.87	.80	.85	.82	.95

* Obtained from spring 2013 reliabilities reported above in the Reliability section.

† Obtained from the *ACT Explore Technical Manual* (ACT, 2013a)

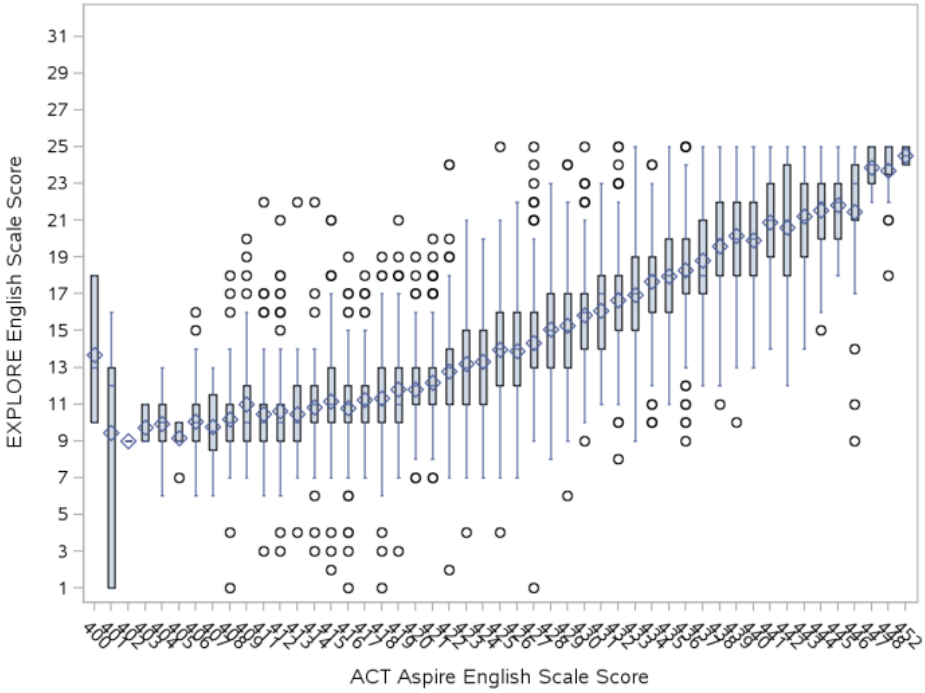
‡ Obtained from the *ACT Plan Technical Manual* (ACT, 2013b)

Disattenuated correlations also provide us with an estimated upper limit of the observed correlations in table 9.2. With this in mind, many of the moderate observed correlations in table 9.2 are not that far from their disattenuated values, suggesting that they couldn't be much higher than those observed due to the reliability of the contributing tests. For example, the smallest observed correlations in table 9.2 were between Reading tests (.70, .66 and .65), but the disattenuated correlations were roughly .10 higher (.83, .76, and .76).

Figures 9.1–9.5 display the relationships between ACT Explore or ACT Plan scale scores and ACT Aspire scale scores using box plots. Each box plot represents the distribution of ACT Explore or ACT Plan scale scores for a particular ACT Aspire scale score, with each box representing the 25th percentile to the 75th percentile of ACT Explore/ACT Plan scores, the line in the middle of the box representing the 50th percentile (or median), and the diamond representing the mean. The upper and lower whiskers of the box plots represent plus or minus 1.5 times the interquartile range (i.e., the range represented by the difference between 25th and 75th percentile), and the circles represent individual scores outside of the range represented by the whiskers. As ACT Aspire scale scores increase (horizontal axis) the boxes representing the ACT Explore or ACT Plan scales score distribution also generally increase (vertical axis), illustrating the positive relationship between scale scores.

Figure 9.1. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire English scale score

Grade 8



Grade 9

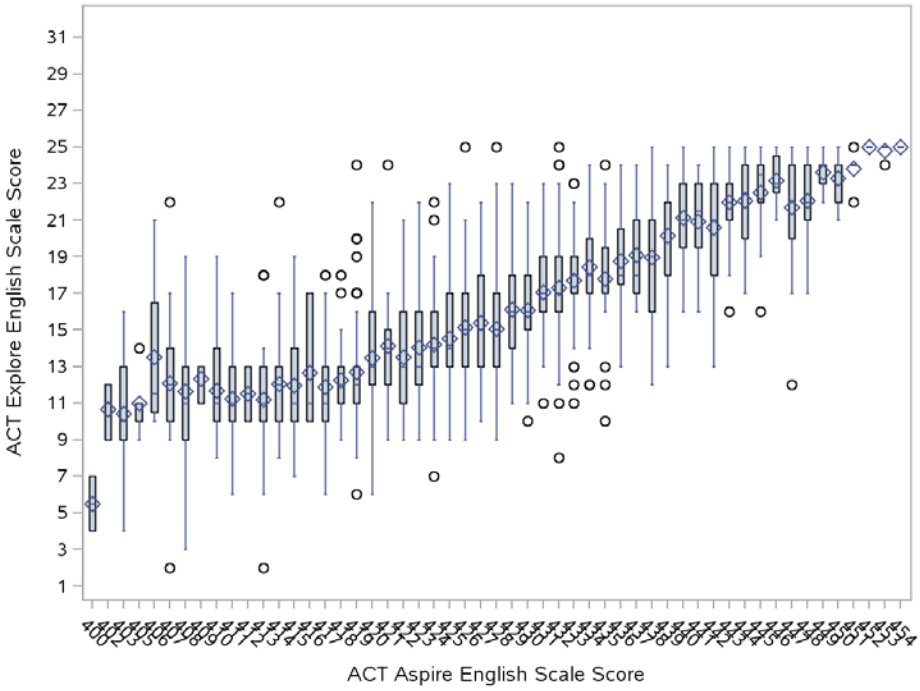


Figure 9.1. (continued)

Grade 10

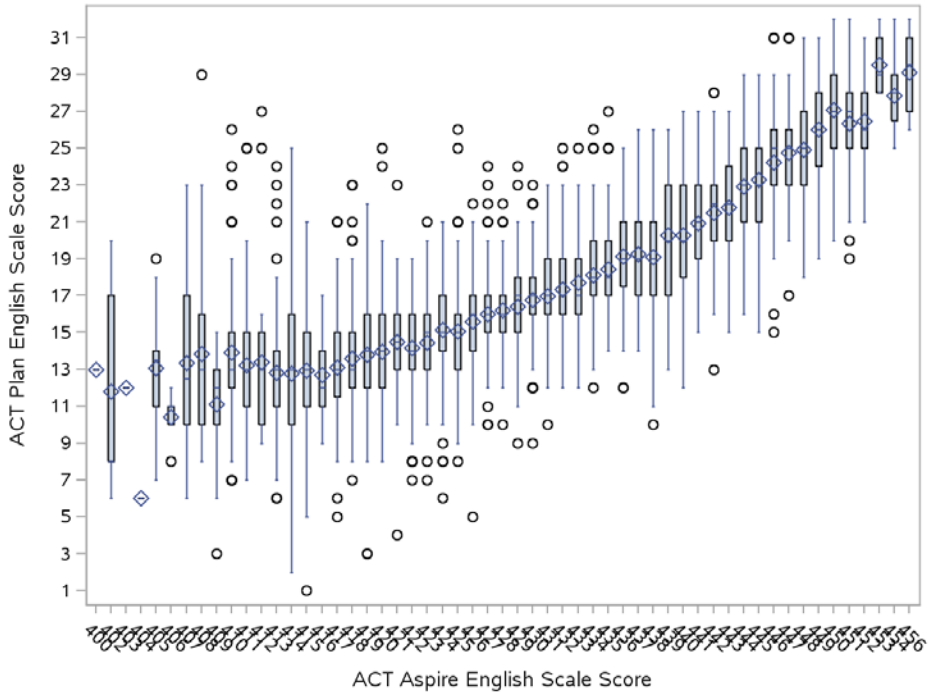


Figure 9.2. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Mathematics scale score

Grade 8

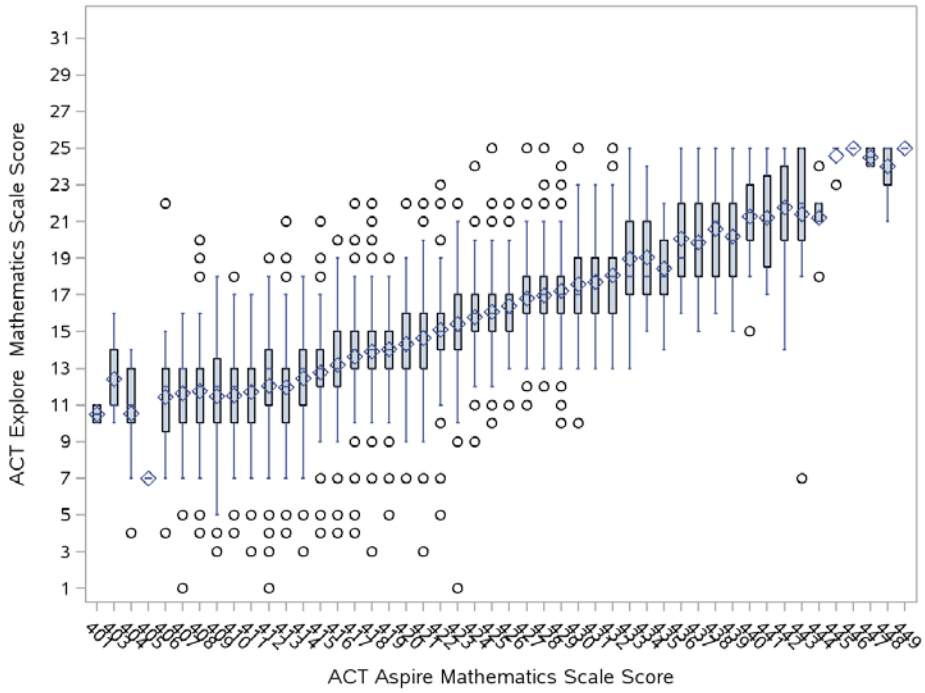
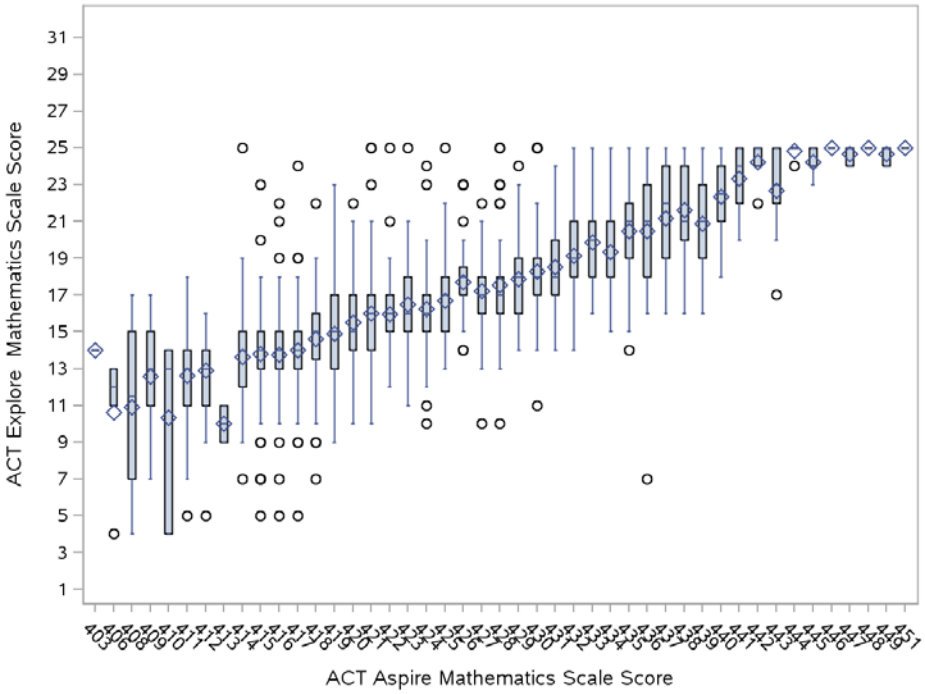


Figure 9.2. (continued)

Grade 9



Grade 10

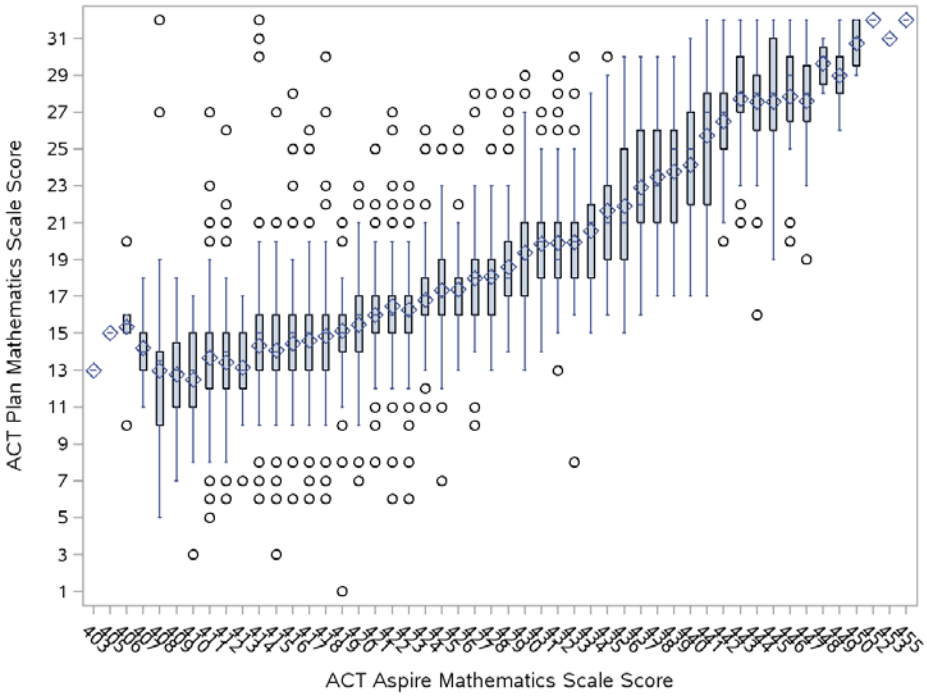
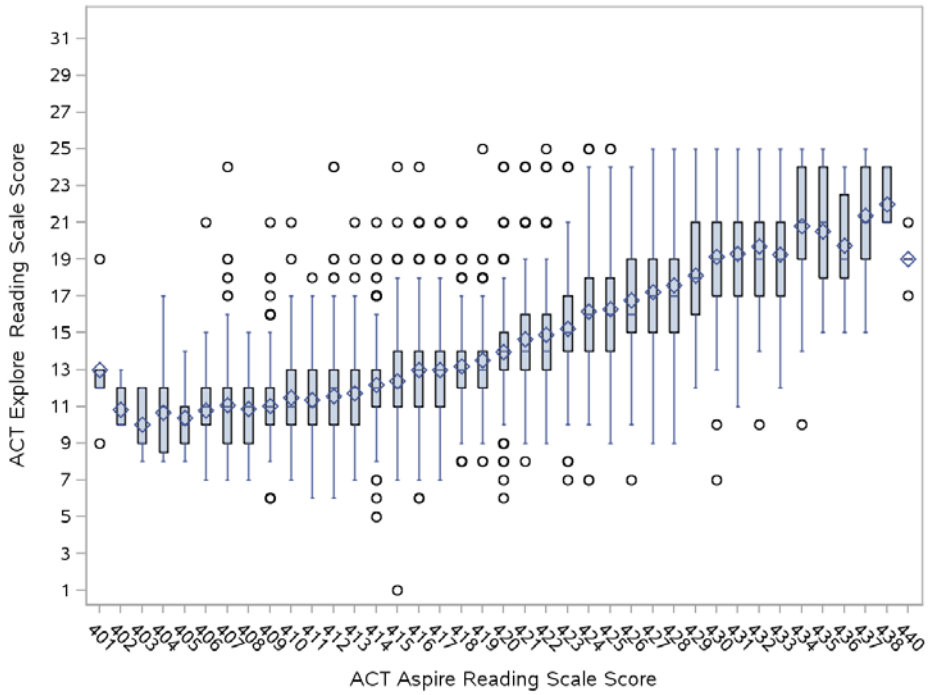


Figure 9.3. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Reading scale score

Grade 8



Grade 9

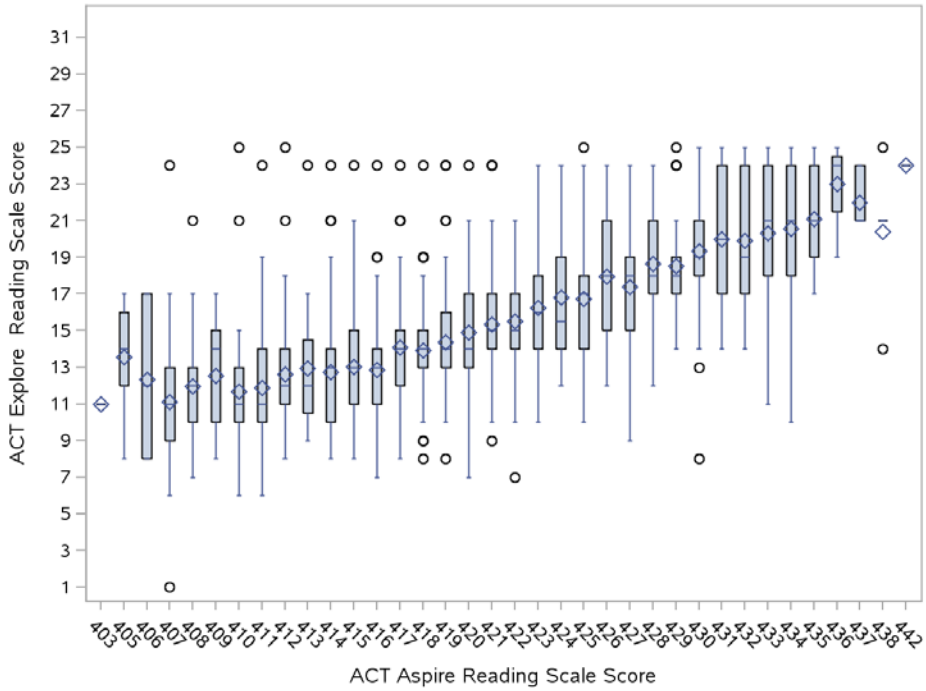


Figure 9.3. (continued)

Grade 10

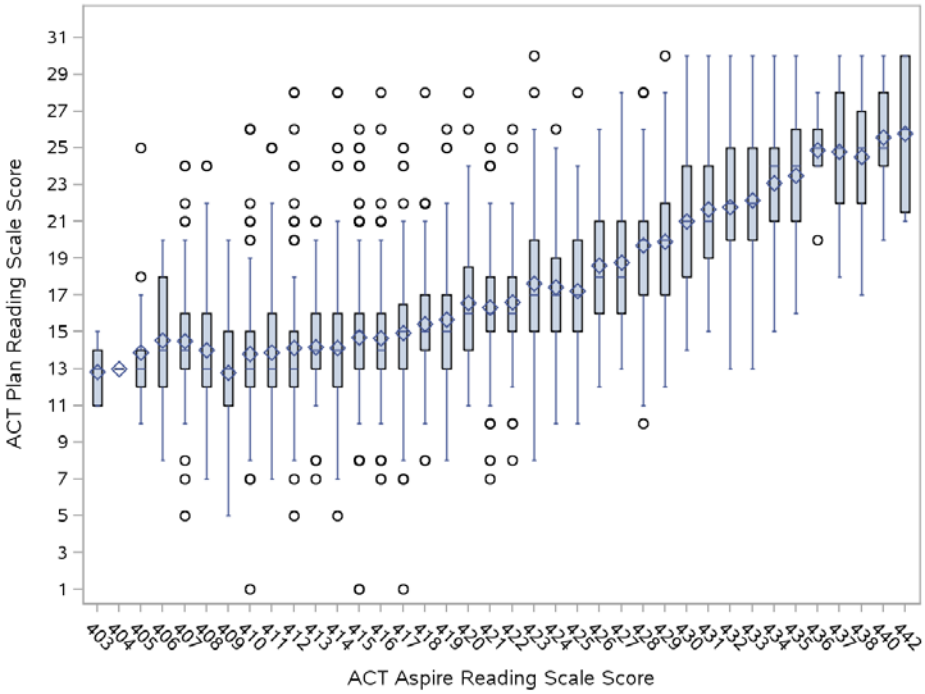


Figure 9.4. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Science scale score

Grade 8

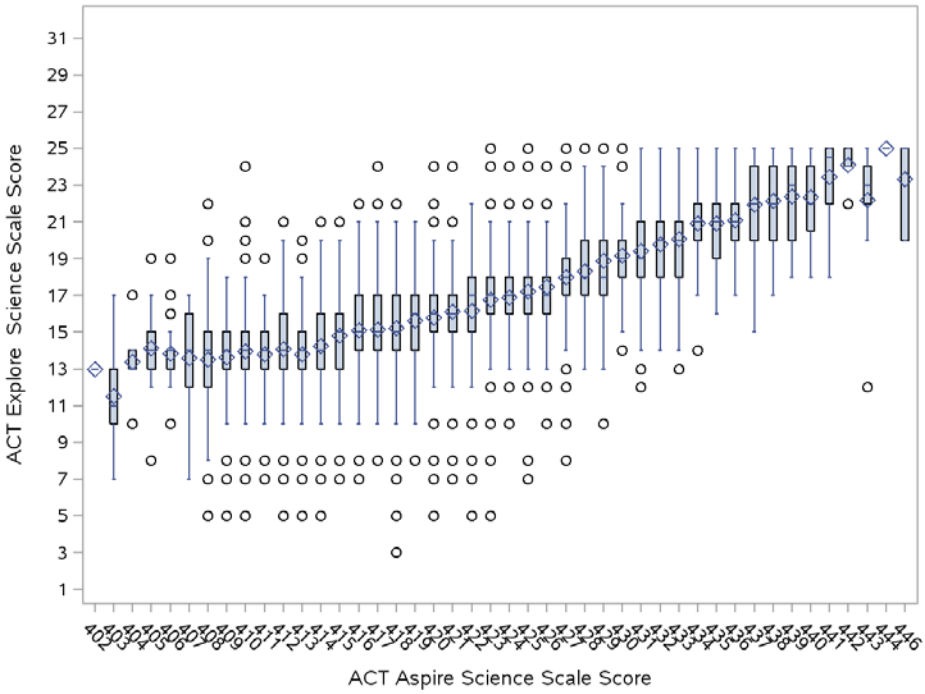
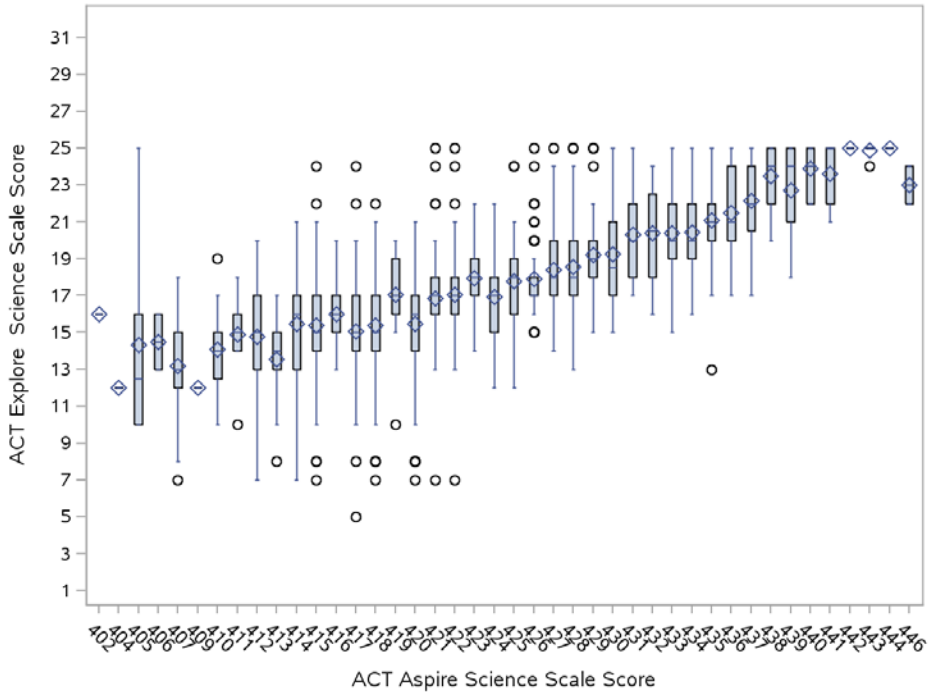


Figure 9.4. (continued)

Grade 9



Grade 10

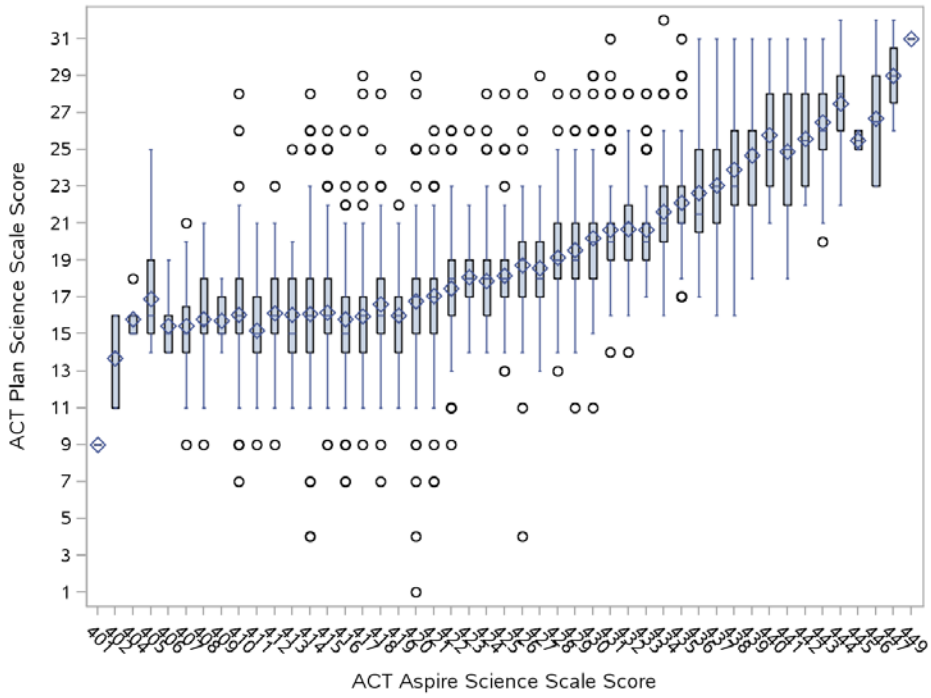
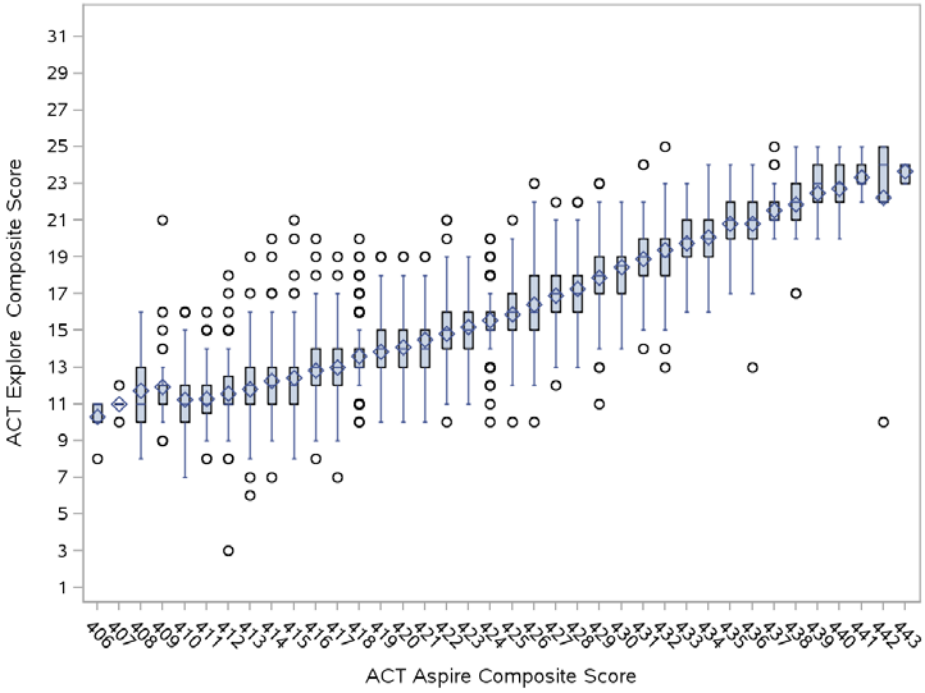


Figure 9.5. Box plots of ACT Explore or ACT Plan scale scores for each ACT Aspire Composite scale score

Grade 8



Grade 9

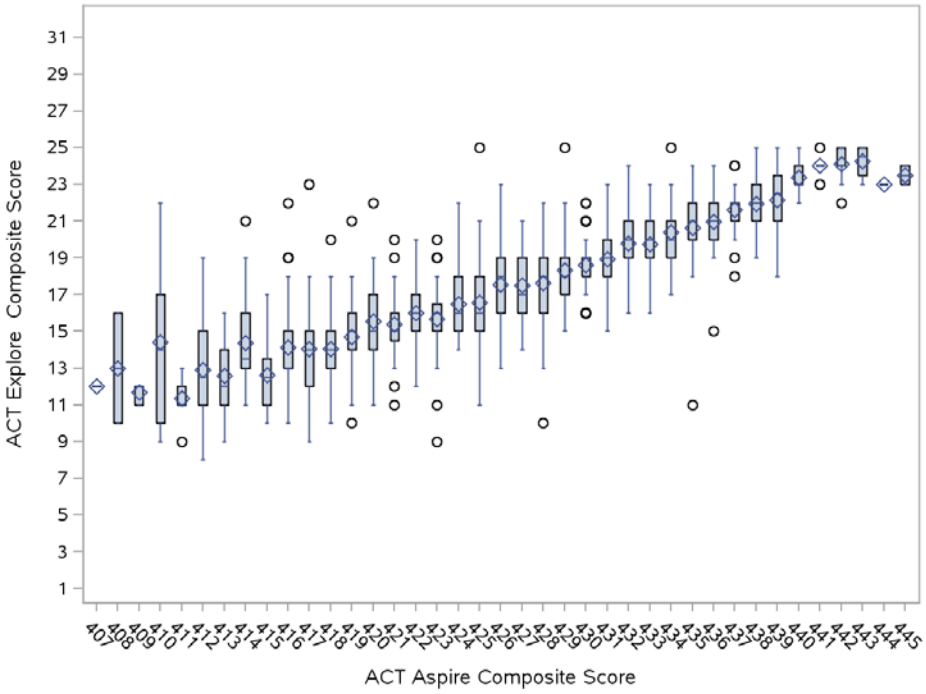
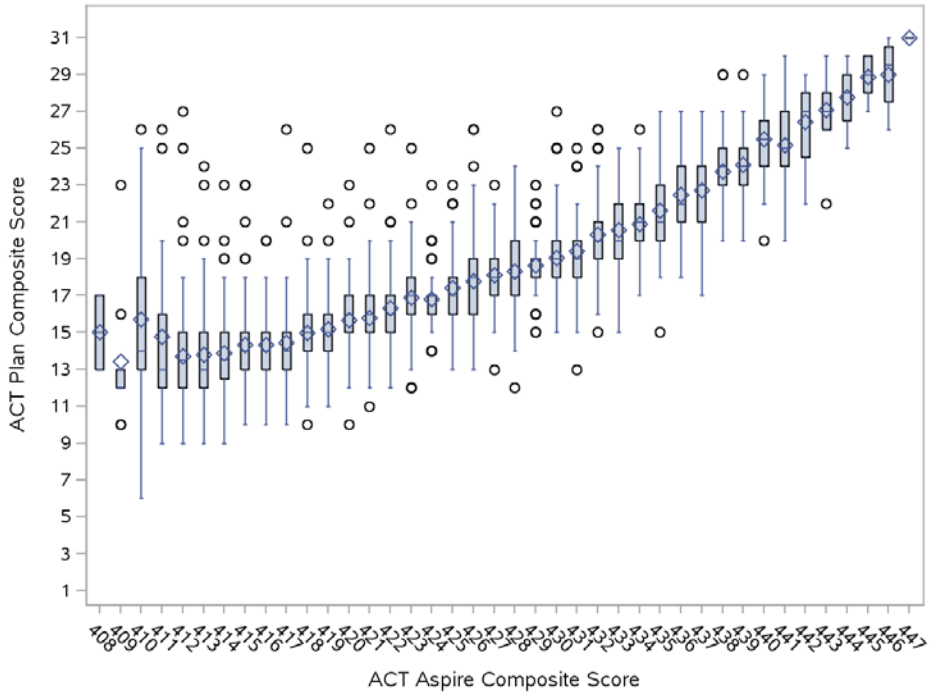


Figure 9.5. (continued)

Grade 10



Together, the correlations and box plots show that ACT Explore/ACT Plan and ACT Aspire scores are moderately to strongly related. Because ACT Explore/ACT Plan and ACT Aspire are designed to measure student achievement and have similar purposes, we would expect to observe relatively strong, but not perfect, positive relationships between scale scores across assessments, similar to those observed in tables 9.2 and 9.3. The observed correlations were lower than anticipated, and this appeared to be explained by the reliabilities of the tests tempering the relationships between scores.

In addition to considering the relationships between scale scores across ACT Explore or ACT Plan and ACT Aspire for the same subject, we can consider the pattern of relationships across subjects (in this case, English, mathematics, reading, and science) and assessments (ACT Explore/ACT Plan and ACT Aspire). This analysis is commonly referred to as investigating a *multitrait-multimethod matrix* (Campbell and Fiske 1959). This matrix is intended to provide convergent and divergent evidence regarding the subjects (traits) being measured by different assessments (methods).

Table 9.5 contains a multitrait-multimethod matrix that includes ACT Explore/ACT Plan and ACT Aspire English, Mathematics, Reading, and Science scores for grades 8–10. Test reliabilities are reported in parentheses when a row and column contains the same subject and same test (for example, ACT Explore Reading with ACT Explore Reading)[12] and the remaining cells contain correlations among scores. Ideally, we

[12] These reliabilities are Cronbach's alpha.

would include a variety of assessments measuring performance in different subjects to study effects due to methods and those due to traits. However, availability of traits and methods was limited to ACT Explore/ACT Plan scores and ACT Aspire English, Reading, Mathematics, and Science scores for this analysis.

Table 9.5. Multitrait-Multimethod Matrices for ACT Explore/Plan and ACT Aspire Scale Scores by Grade Level

		ACT Explore/ACT Plan				ACT Aspire			
Grade (valid *N*)	**Scale Score**	**English**	**Math**	**Reading**	**Science**	**English**	**Math**	**Reading**	**Science**
Grade 8 (6,419)	ACT Explore English	(.84)							
	ACT Explore Math	.68	(.76)						
	ACT Explore Reading	.76	.63	(.86)					
	ACT Explore Science	.70	.67	.71	(.79)				
	ACT Aspire English	**.76**	.62	.69	.63	(.85)			
	ACT Aspire Math	.69	**.72**	.63	.66	.71	(.88)		
	ACT Aspire Reading	.68	.57	**.70**	.63	.73	.67	(.85)	
	ACT Aspire Science	.69	.64	.68	**.69**	.72	.77	.76	(.87)
Grade 9 (1,181)	ACT Explore English	(.86)							
	ACT Explore Math	.70	(.80)						
	ACT Explore Reading	.75	.66	(.86)					
	ACT Explore Science	.69	.70	.73	(.82)				
	ACT Aspire English	**.75**	.62	.70	.64	(.91)			
	ACT Aspire Math	.64	**.73**	.60	.67	.70	(.89)		
	ACT Aspire Reading	.62	.55	**.65**	.60	.75	.67	(.88)	
	ACT Aspire Science	.64	.64	.63	**.68**	.73	.78	.75	(.88)
Grade 10 (3,154)	ACT Plan English	(.87)							
	ACT Plan Math	.73	(.80)						
	ACT Plan Reading	.78	.66	(.85)					
	ACT Plan Science	.74	.76	.71	(.82)				
	ACT Aspire English	**.77**	.63	.68	.68	(.91)			
	ACT Aspire Math	.69	**.77**	.62	.73	.74	(.89)		
	ACT Aspire Reading	.67	.57	**.65**	.64	.76	.69	(.88)	
	ACT Aspire Science	.66	.67	.63	**.71**	.73	.78	.75	(.88)

Note: Cells in parentheses are Cronbach's alpha reliabilities. Others are correlations.

For convergent evidence that ACT Aspire and ACT Explore/ACT Plan are measuring English, reading, mathematics, and science achievement, we want to see large positive *monotrait-heteromethod* correlations, which are represented by the correlation between ACT Explore/ACT Plan and ACT Aspire in table 9.5 (see bold text). In addition, we want to see smaller positive *heterotrait-monomethod* correlations, which are represented by all of the correlations between subjects within a given test (ACT Explore or ACT Plan or ACT Aspire) in table 9.5 (e.g., correlations between ACT Explore in English, mathematics, reading, and science). Finally, we want to see the smallest correlations for *heterotrait-heteromethod* correlations, which are represented by the correlations between different subjects across different tests. To support interpreting ACT Explore/ACT Plan and ACT Aspire test scores as measuring distinct academic achievement in a subject, we want to see stronger convergent evidence (monotrait-heteromethod correlations) and weaker discriminant evidence (heterotrait-heteromethod and heterotrait-monomethod correlations).

For English, the monotrait-heteromethod correlations (ACT Explore/ACT Plan English and ACT Aspire English) were all relatively large (above .75) and were among the largest in table 9.5. Within ACT Aspire, the heterotrait-monomethod correlations for English were also relatively large, although none were larger than the monotrait-heteromethod correlations. Within ACT Explore and ACT Plan, the heterotrait-monomethod correlations for English were relatively large, and for Grades 8 to 10 Reading the correlations were the same as or larger than the monotrait-heteromethod correlations. Most of the heterotrait-heteromethod correlations were the lowest but were not that much lower than the heterotrait-monomethod correlations. The English test showed slightly stronger convergent evidence for ACT Aspire than for ACT Explore/ACT Plan, but the large heterotrait-monomethod correlations indicate that the English test was not strongly differentiated from the other subjects.

For Mathematics, the monotrait-heteromethod correlations ranged from .72 to .77. Within ACT Aspire, the heterotrait-monomethod correlations for Mathematics were moderate to large, and the correlations between ACT Aspire Science and ACT Aspire Mathematics were slightly larger (.77 to .78) than the correlation between ACT Explore/ACT Plan and ACT Aspire Mathematics (.72 to .77). Within ACT Explore/ACT Plan the heterotrait-monomethod correlations for Mathematics were moderate to large but all were smaller than the monotrait-heteromethod correlations. The heterotrait-heteromethod correlations were smaller than the other correlations. The Mathematics test showed slightly stronger convergent evidence for ACT Explore/ACT Plan than for ACT Aspire; ACT Aspire Mathematics and Science were most closely related.

For Reading, the monotrait-heteromethod correlations ranged from .65 to .70. For ACT Explore/ACT Plan and ACT Aspire, the heterotrait-monomethod correlations between Reading and English and between Reading and Science were larger than the monotrait-heteromethod correlations. In addition, for ACT Aspire, the correlation between Reading and Mathematics in grade 10 was larger than the correlation between Reading Tests for ACT Plan and ACT Aspire. The heterotrait-heteromethod correlations for Grade 10 English and Reading were larger than the monotrait-heteromethod correlations for Grade 10 Reading. The ACT Aspire and ACT Explore/ACT Plan Reading Tests showed similar patterns of relatively weak convergent evidence compared to discriminant evidence; the Reading tests showed particularly strong method effects.

For Science, the monotrait-heteromethod correlations ranged from .68 to .71. Nearly all of the heterotrait-monomethod correlations were larger than the monotrait-heteromethod correlations. The one exception was the grade 8 ACT Explore Mathematics and Science correlation, which was .67. The heterotrait-heteromethod correlations were smaller than the heterotrait-monomethod correlations, although for grade 8 the ACT Aspire Science and ACT Explore English correlation was the same as the monotrait-heteromethod correlation for science (.69), and for grade 10 the ACT Plan Science and ACT Aspire Mathematics correlation was larger than the monotrait-heteromethod correlation for Science (which was .71).

The multitrait-multimethod matrix for ACT Explore/ACT Plan and ACT Aspire provided evidence regarding how well the subjects (traits) were differentiated from the assessments (methods). This evidence was mixed. English showed the strongest convergent evidence, particularly for ACT Aspire, followed by Mathematics, Reading, and Science. ACT Aspire Mathematics and Science showed particularly strong relationships. While ACT Explore/ACT Plan and ACT Aspire are intended to measure the same traits, the multitrait-multimethod results are consistent with the argument that the assessments may be systematically different in ways that lead to stronger relationships among scores within each assessment compared to between the assessments, particularly for Reading and Science. One explanation is that ACT Aspire forms contain a variety of item types, including technology-enhanced and fill-in-the-blank items (online) and constructed-response items. Interestingly, the English test, which showed the strongest convergent evidence, does not have constructed-response items.

While there may be some degree of method effects, the results of this study support the argument that ACT Aspire is measuring achievement in English, mathematics, reading, and science in grades 8, 9, and 10. The correlations of ACT Aspire scores with ACT Explore and ACT Plan, tests that are intended to measure achievement and have been separately validated, are moderate to large. The magnitudes of these correlations are similar to those observed between ACT Explore and ACT Plan and between ACT Explore and The ACT in other technical documentation (e.g., ACT 2007b, 84).

Study 2: Comparison of State Assessment Scores to ACT Aspire Scores

In this study the relationships between Alabama Reading and Mathematics Test (plus Science; ARMT+) and ACT Aspire scale scores were compared using a sample of Alabama students taking both assessments in spring of 2013 in grades 3–8.[13] This sample included students participating in 2013 ACT Aspire special studies described in other chapters. The sample sizes by grade and subject are listed in table 9.6 and included roughly 5% to 15% of the tested population in Alabama.

ARMT+ and ACT Aspire are intended to assess student achievement. ARMT+ was designed to assess students' mastery of state content standards in reading, mathematics, and science. ACT Aspire is designed to measure students' academic achievement in English, mathematics, reading, science, and writing. While each test is built using different blueprints and different specifications the traits intended to be measured by both include mathematics, reading, and science.

Table 9.7 lists the means, standard deviations, minimum, and maximum for ARMT+ and ACT Aspire scale scores by grade and subject. In all but three cases, mean scale scores increase across grades. For grades 6 and 7, ACT Aspire Reading mean scale scores decrease slightly from 418.46 to 418.08.[14] For grades 5 and 6 mathematics, ARMT+ mean scale scores decrease from 676.56 to 667.36 and for grades 5 and 7 science, ARMT+ mean scale scores decrease from 596.87 to 563.94. However, scores on ARMT+ are not strictly interpreted assuming vertical scale properties. For this reason, grade-to-grade comparisons of ARMT+ and ACT Aspire scores with data from 2013 administrations were not entirely appropriate. In this study we compared ARMT+ scores and ACT Aspire scores within grades, although many tables and figures include multiple grade levels to save space.

[13] Alabama science scores were only available for grades 5 and 7.

[14] This decrease in mean scale scores is consistent with results from other samples of students, including those included in the scaling study for ACT Aspire.

Table 9.6. Sample Sizes by Grade and Subject for the Sample of Students with Scores on ARMT+ and ACT Aspire in Spring 2013

Grade	Mathematics	Reading	Science	Total
3	9,020	9,198	—	9,688
4	9,185	9,249	—	9,849
5	8,188	8,231	7,209	9,011
6	6,257	6,242	—	6,695
7	5,060	5,058	4,702	5,423
8	4,045	4,306	—	4,475
Total	41,755	42,284	11,911	45,141

Table 9.7. Descriptive Statistics for ARMT+ and ACT Aspire Scale Scores

Subject	Grade	N	Assessment	Mean	SD	Minimum	Maximum
Mathematics	3	9,020	Alabama	635.27	39.77	503	781
			ACT Aspire	412.18	3.91	400	429
	4	9,185	Alabama	649.85	44.12	471	865
			ACT Aspire	414.91	3.87	402	434
	5	8,188	Alabama	676.56	38.18	560	846
			ACT Aspire	416.85	4.78	402	437
	6	6,257	Alabama	667.36	34.28	563	823
			ACT Aspire	418.12	5.53	402	445
	7	5,060	Alabama	678.25	42.23	573	873
			ACT Aspire	419.10	6.59	401	445
	8	4,045	Alabama	696.32	33.05	606	835
			ACT Aspire	420.21	7.58	401	449
Reading	3	9,198	Alabama	634.60	37.59	471	785
			ACT Aspire	412.28	5.24	401	429
	4	9,249	Alabama	651.80	37.70	513	789
			ACT Aspire	414.60	5.64	401	431
	5	8,231	Alabama	663.56	35.77	549	849
			ACT Aspire	416.45	6.19	401	434
	6	6,242	Alabama	671.91	35.70	538	802
			ACT Aspire	418.46	6.66	401	436
	7	5,058	Alabama	681.49	33.33	554	787
			ACT Aspire	418.08	6.69	402	438
	8	4,306	Alabama	681.71	31.50	577	818
			ACT Aspire	420.32	7.19	401	440
Science	5	7,209	Alabama	596.87	103.50	136	999
			ACT Aspire	418.31	6.43	401	438
	7	4,702	Alabama	563.94	106.92	42	999
			ACT Aspire	418.58	7.42	401	440

Table 9.8 lists the correlations between same-subject ARMT+ scale scores and ACT Aspire scale scores across grades. Correlations ranged from .60 to .81, with higher correlations observed as grade increased for mathematics and science. These correlations are indicative of moderate to strong linear relationships between ARMT+ and ACT Aspire scores across subjects. From a linear regression perspective, we can say that 36% to 66% of the variance in scale scores is shared between ARMT+ and ACT Aspire.[15]

Table 9.9 lists the disattenuated correlations between ARMT+ and ACT Aspire. Published or available reliability coefficients for ARMT+ and ACT Aspire were used to calculate disattenuated correlations, which ranged from .71 to .90 and are indicative of moderate to strong correlations. From a linear regression perspective, we can say that 50% to 81% of the variance in true scores is shared between ARMT+ and ACT Aspire.

Table 9.8. Correlations between ARMT+ Scale Scores and ACT Aspire Scale Scores

Grade	Mathematics	Reading	Science
3	.60	.73	—
4	.62	.77	—
5	.64	.76	.65
6	.64	.74	—
7	.72	.72	.69
8	.81	.75	—

Table 9.9. Disattenuated Correlations between ARMT+ Scale Scores and ACT Aspire Scale Scores

Grade	Mathematics	Reading	Science
3	.71	.84	—
4	.77	.88	—
5	.79	.87	.73
6	.77	.85	—
7	.84	.84	.76
8	.90	.85	—

[15] Percentage of variance shared is calculated by squaring the correlation between scores.

Figures 9.6–9.8 display the relationships between ARMT+ scale scores and ACT Aspire scale scores using box plots. Each box plot represents the distribution of ARMT+ scale scores for a particular ACT Aspire scale score. The important pattern to observe in figures 9.6–9.8 is that as ACT Aspire scale scores increase (horizontal axis) the boxes representing the ARMT+ scales score distribution also generally increase (vertical axis).

Together, the correlations and box plots show that ARMT+ and ACT Aspire scores are moderately to strongly related. Because ARMT+ and ACT Aspire are both designed to measure student achievement, but each are distinct assessments built to different test specifications with their own unique scale, we would expect to observe moderate to strong, but not perfect, positive relationships between ARMT+ scale scores and ACT Aspire scale scores.

Table 9.10 contains a multitrait-multimethod matrix which includes scores in mathematics, reading, and science for ARMT+ and ACT Aspire. Test reliabilities are reported in parentheses when a row and column contains the same subject and same test (for example, ARMT+ reading with ARMT+ reading) and the remaining cells contain correlations between scores. Ideally, we would include a variety of assessments measuring performance in different subjects to study effects due to methods and due to traits. However, availability of methods and traits was limited to scores in reading, mathematics, and science for ARMT+ and ACT Aspire in this analysis.

Figure 9.6. Box plots of ARMT+ mathematics scale scores for each ACT Aspire Mathematics scale score

Grade 3

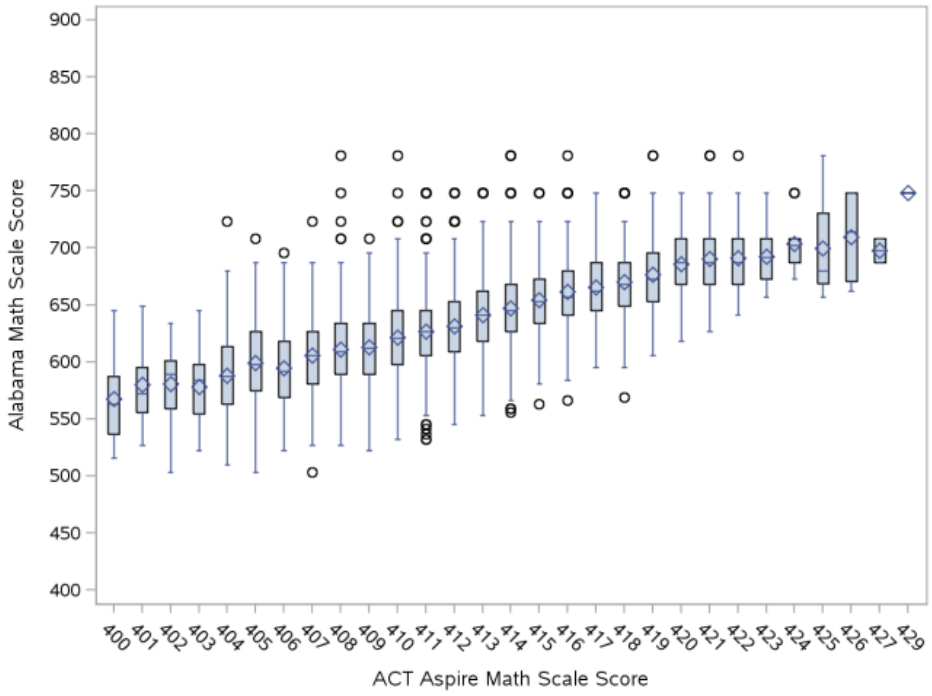
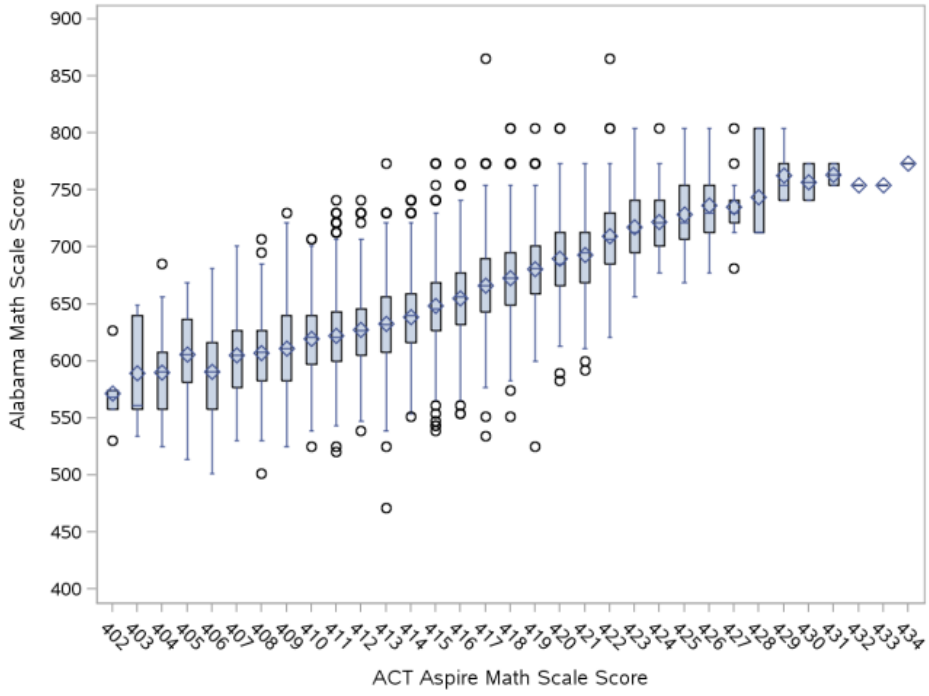


Figure 9.6. (continued)

Grade 4



Grade 5

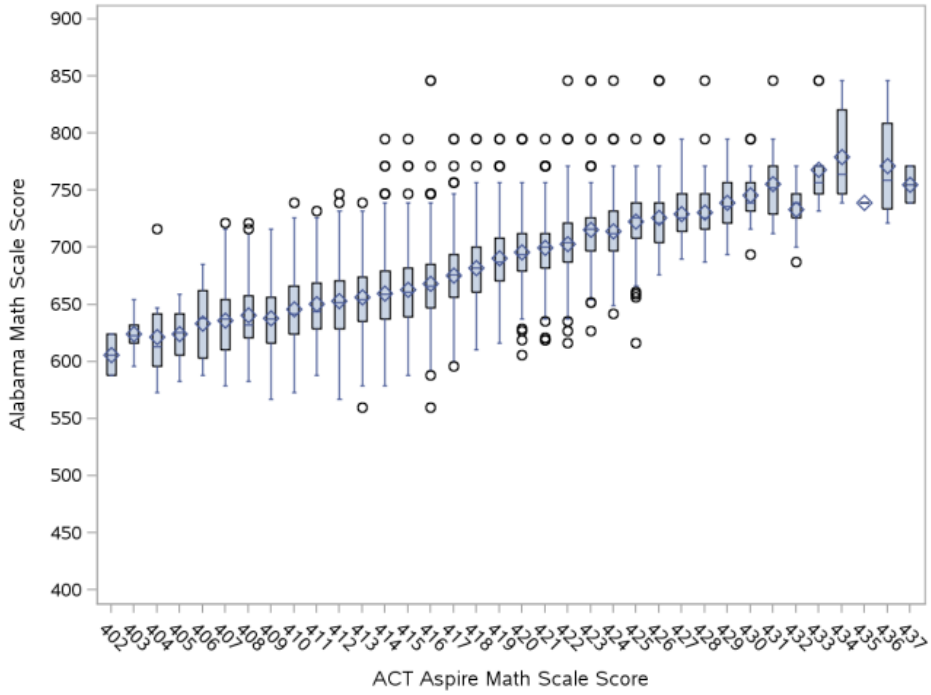
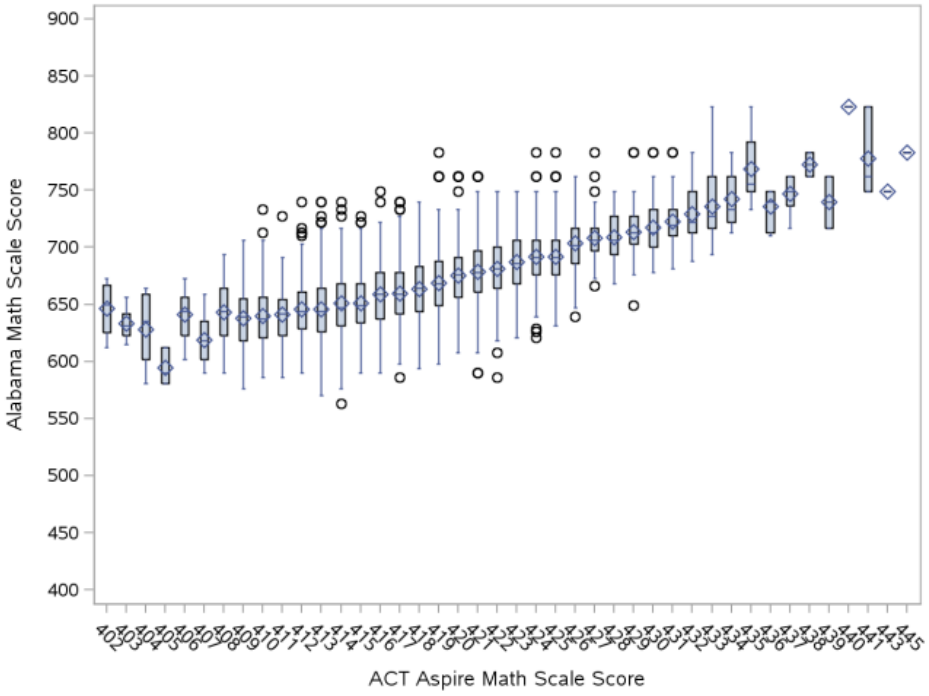


Figure 9.6. (continued)

Grade 6



Grade 7

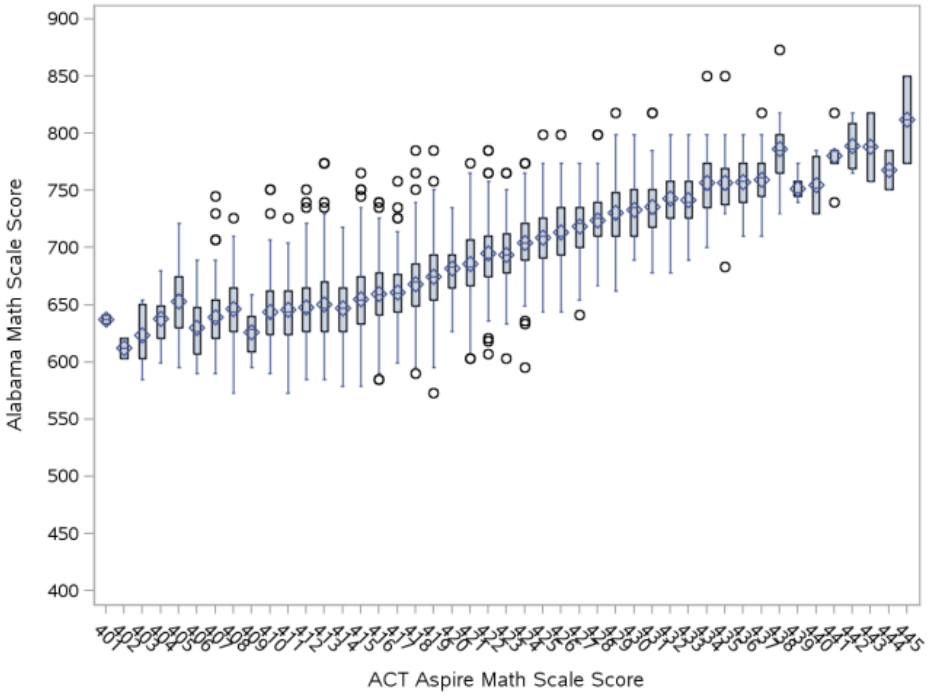


Figure 9.6. (continued)

Grade 8

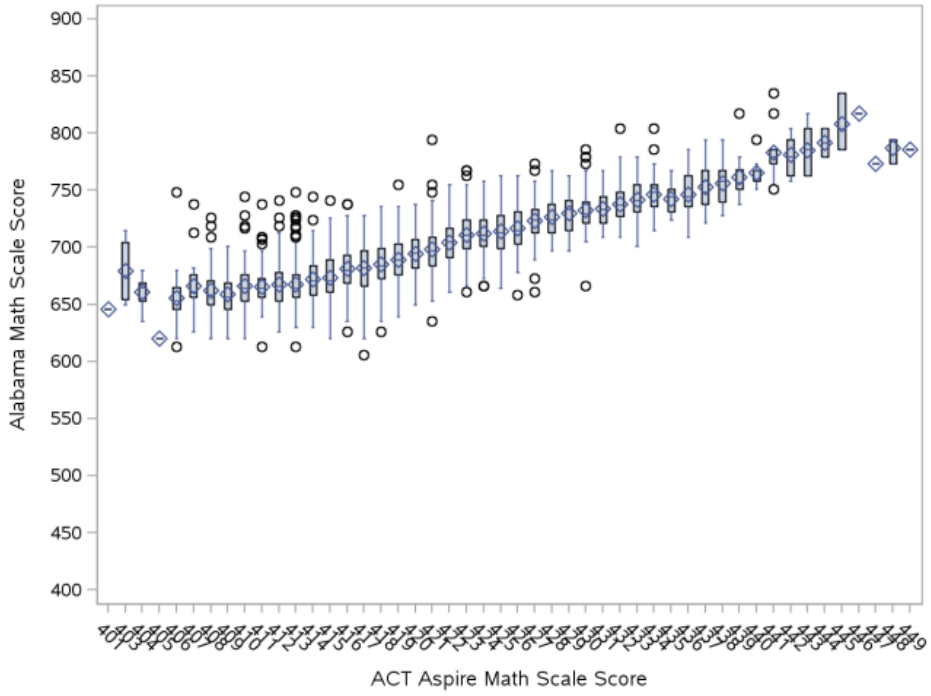


Figure 9.7. Box plots of ARMT+ reading scale scores for each ACT Aspire Reading scale score

Grade 3

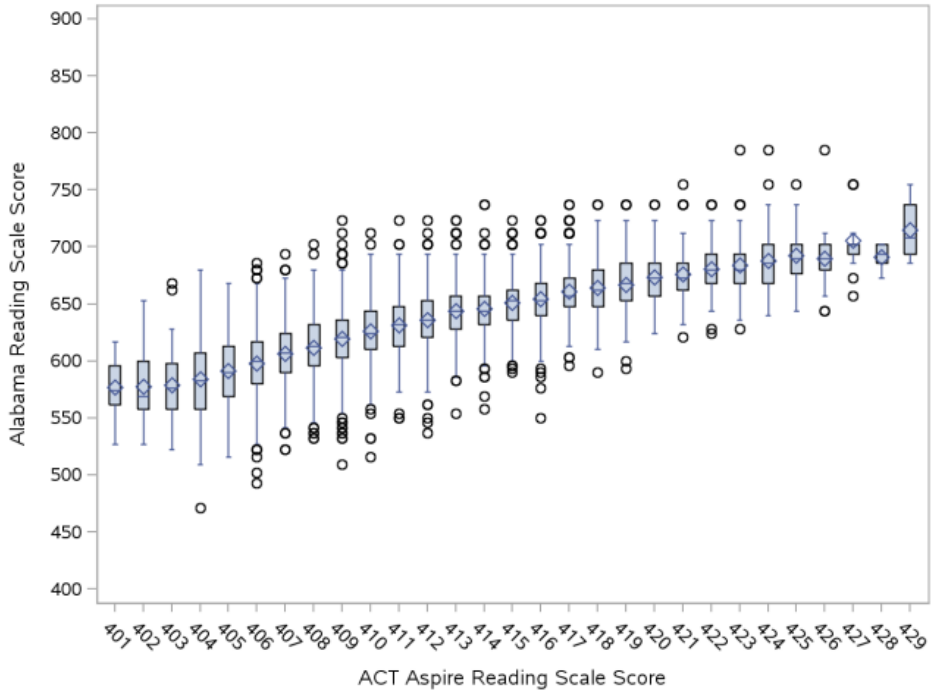
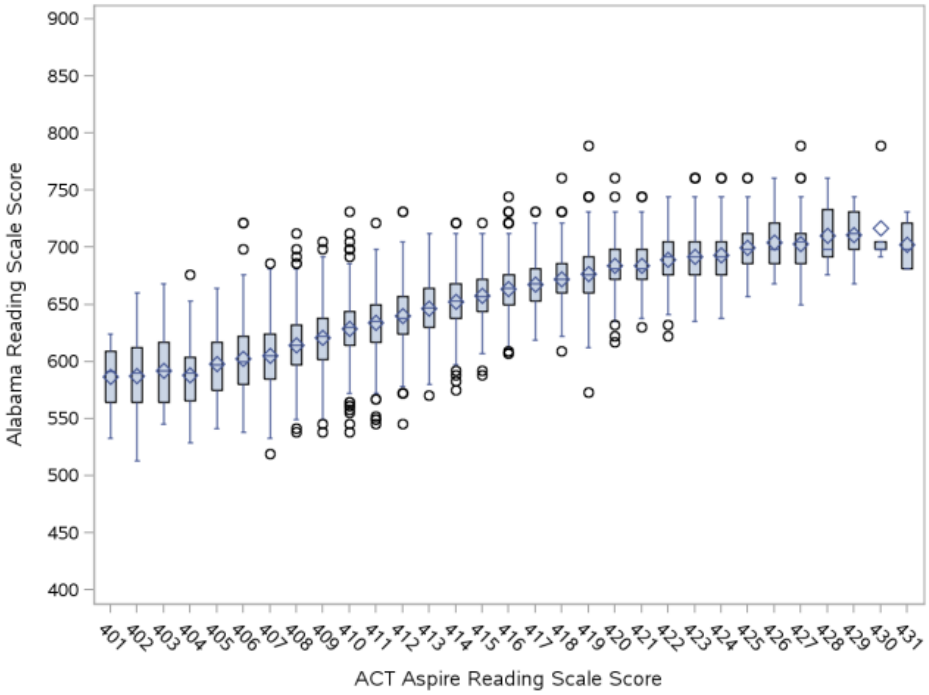


Figure 9.7. (continued)

Grade 4



Grade 5

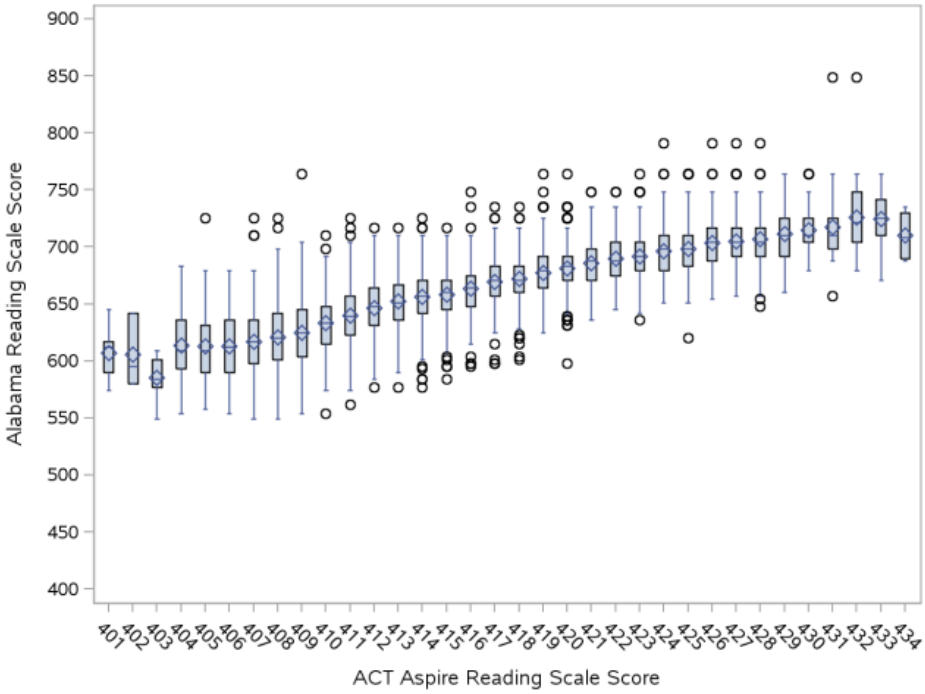
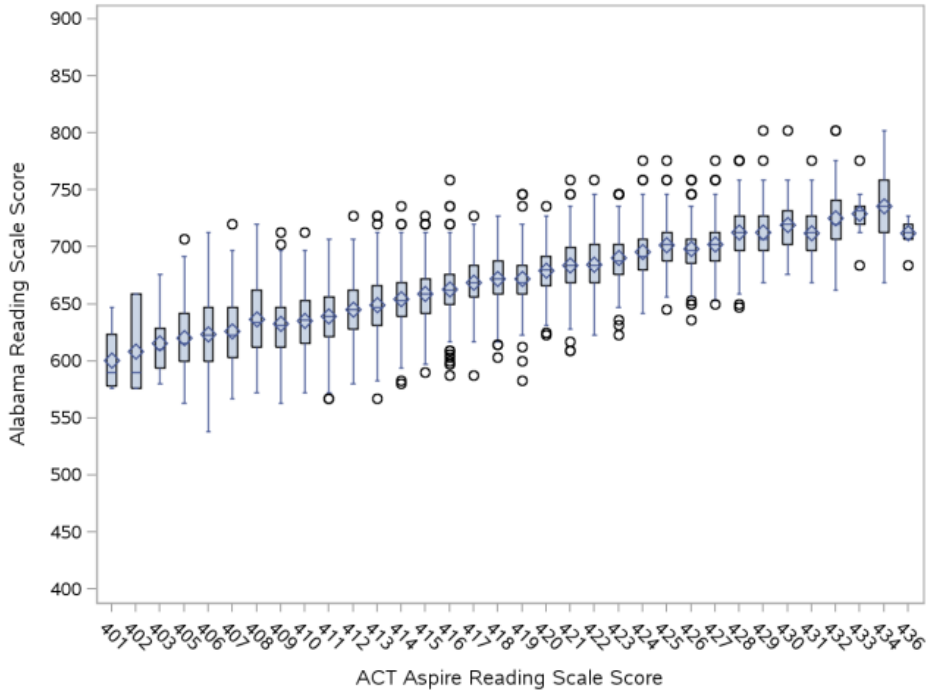


Figure 9.7. (continued)

Grade 6



Grade 7

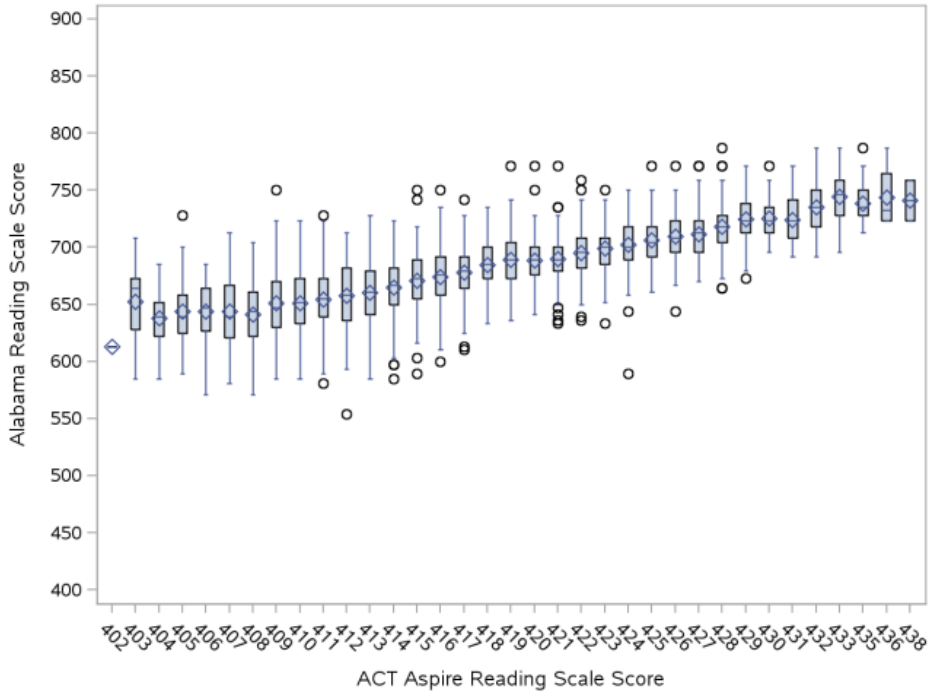


Figure 9.7. (continued)

Grade 8

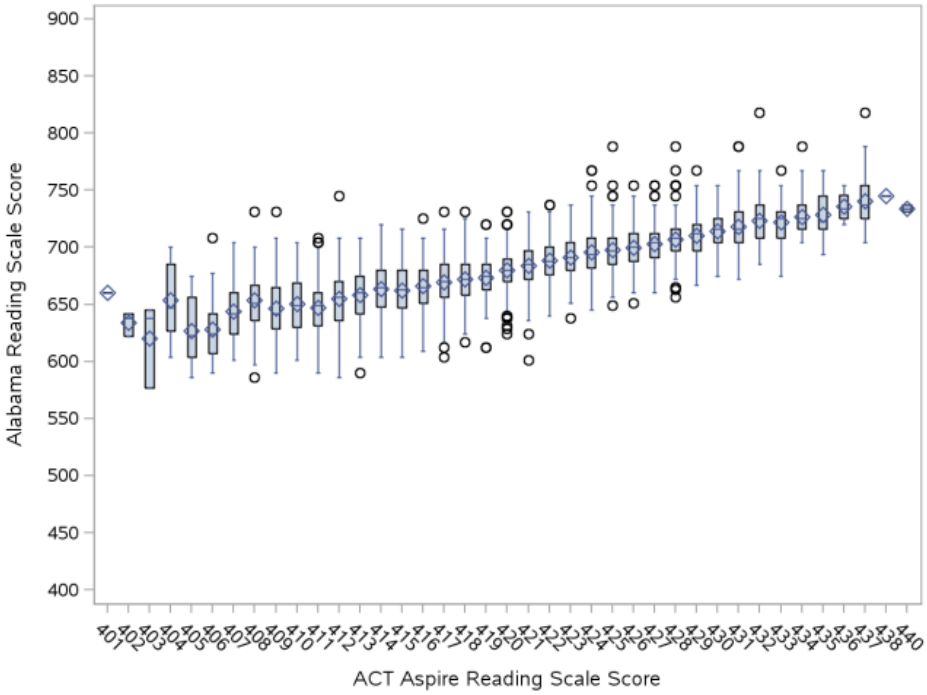


Figure 9.8. Box plots of ARMT+ science scale scores for each ACT Aspire Science scale score

Grade 5

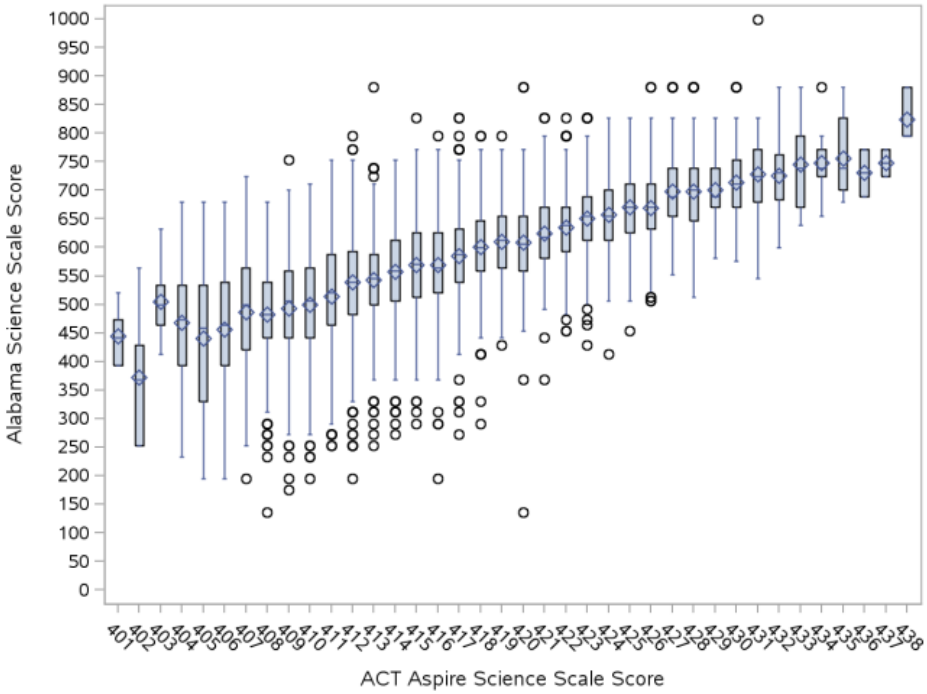


Figure 9.8. (continued)

Grade 7

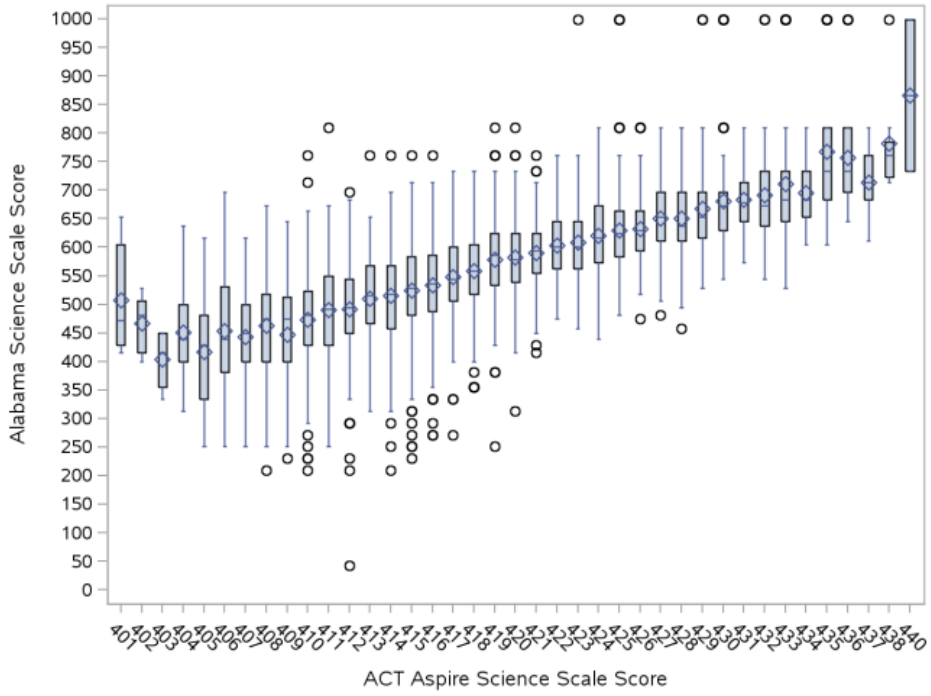


Table 9.10. Multitrait-Multimethod Matrices for ARMT+ and ACT Aspire Scale Scores by Grade Level

		Alabama			ACT Aspire		
Grade (valid N)	**Scale Score**	**Math**	**Reading**	**Science**	**Math**	**Reading**	**Science**
Grade 3 (8,530)	Alabama Math	(.92)					
	Alabama Reading	.71	(.90)				
	ACT Aspire Math	**.60**	.63		(.77)		
	ACT Aspire Reading	.56	**.73**		.65	(.84)	
Grade 4 (8,585)	Alabama Math	(.93)					
	Alabama Reading	.71	(.92)				
	ACT Aspire Math	**.62**	.57		(.69)		
	ACT Aspire Reading	.62	**.77**		.57	(.84)	
Grade 5 (6,543)	Alabama Math	(.92)					
	Alabama Reading	.73	(.91)				
	Alabama Science	.70	.73	(.92)			
	ACT Aspire Math	**.63**	.57	.51	(.71)		
	ACT Aspire Reading	.65	**.76**	.63	.58	(.83)	
	ACT Aspire Science	.70	.73	**.65**	.64	.75	(.86)
Grade 6 (5,804)	Alabama Math	(.92)					
	Alabama Reading	.73	(.91)				
	ACT Aspire Math	**.63**	.53		(.76)		
	ACT Aspire Reading	.64	**.74**		.54	(.84)	
Grade 7 (4,345)	Alabama Math	(.94)					
	Alabama Reading	.73	(.90)				
	Alabama Science	.75	.74	(.94)			
	ACT Aspire Math	**.71**	.59	.60	(.78)		
	ACT Aspire Reading	.68	**.72**	.69	.63	(.82)	
	ACT Aspire Science	.73	.70	**.69**	.69	.76	(.88)
Grade 8 (3,876)	Alabama Math	(.93)					
	Alabama Reading	.75	(.92)				
	ACT Aspire Math	**.82**	.70		(.88)		
	ACT Aspire Reading	.64	**.75**		.67	(.85)	

Note: Cells in parentheses are Cronbach's alpha reliabilities. Others are correlations.

For convergent evidence that ACT Aspire and ARMT+ are measuring reading, mathematics, and science achievement, we want to see large positive *monotrait-heteromethod* correlations, which are represented by the correlation between ARMT+ and ACT Aspire scores in the same subject (e.g., ARMT+ reading and ACT Aspire Reading) (see bold type in table 9.10). In addition, we want to see smaller positive *heterotrait-monomethod* correlations, which are represented by all of the correlations among subjects within a given test (ARMT+ or ACT Aspire) in table 9.10 (e.g., correlations among ARMT+ mathematics, reading, and science scores). Finally, we want to see the smallest correlations for *heterotrait-heteromethod* correlations, which are represented by the correlations among different subjects across different tests. To support interpreting ARMT+ and ACT Aspire test scores as measuring distinct academic achievement in a subject, we want to see stronger convergent evidence (monotrait-heteromethod correlations) and weaker discriminant evidence (heterotrait-heteromethod and heterotrait-monomethod correlations).

In reading, we see that the monotrait-heteromethod correlations between ARMT+ and ACT Aspire are relatively large (all are above .7) and for all but grades 7 and 8 these are the largest correlations in the matrices. Moreover, almost all of the heterotrait-monomethod correlations are smaller than the monotrait-heteromethod correlations. The heterotrait-heteromethod correlations are smallest. One anomaly in reading is that for grade 7, the correlation between reading test scores on ARMT+ and ACT Aspire is not as large as the correlation between (a) ARMT+ reading and ARMT+ science and (b) ACT Aspire Reading and ACT Aspire Science. In other words, the method effect is stronger for Grade 7 Reading than the trait effect, at least when considering reading and science. In general, though, these patterns of correlations indicate stronger convergent evidence than discriminant evidence across reading tests for ARMT+ and ACT Aspire.

In science, we see that the monotrait-heteromethod correlations between ARMT+ and ACT Aspire are not as large as many of the heterotrait-monomethod correlations. For example, for grade 5, the science correlation between ARMT+ and ACT Aspire is .65, all of the heterotrait-monomethod correlations for ARMT+ (i.e., those in the first three rows for grade 5) are .7 or higher, and the correlation between ACT Aspire Reading and Science is .75. In science, we see weaker convergent evidence than discriminant evidence, particularly for ARMT+. However, these results may not be surprising given that science is likely to draw on both reading and mathematics.

In mathematics, we see some divergent evidence for ARMT+ in all grades except grade 8. For example, in each of grades 3–7, ARMT+ mathematics correlates more highly with ARMT+ reading than with ACT Aspire Mathematics (heterotrait-monomethod). For grades 5 and 7, ARMT+ mathematics correlates more highly with ARMT+ science than with ACT Aspire Mathematics. In addition, in some cases ARMT+ mathematics correlates more highly with ACT Aspire Reading and Science than with ACT Aspire Mathematics (heterotrait-heteromethod). ACT Aspire

Mathematics scores also show some divergent evidence. For example, ACT Aspire Grade 3 Mathematics scores are more highly correlated with ACT Aspire Reading scores than with ARMT+ mathematics scores, and ACT Aspire Grade 5 Mathematics scores have slightly higher correlations with ACT Aspire Science scores than with ARMT+ mathematics scores.

One possible explanation for the divergent evidence observed in science and mathematics is the effects of the test reliabilities on the correlations among variables. Table 9.11 lists the disattenuated correlations among ARMT+ and ACT Aspire scale scores. As mentioned earlier, these correlations are an estimate of the *true score* relationships after taking into account the reliabilities of the two scores. If we consider the disattenuated correlations as an upper limit on the observable correlations between scores (i.e., those in table 9.10), then the correlations observed in table 9.9 are quite strong. For example, the correlation between ACT Aspire Grade 4 Mathematics and reading was .57 with a disattenuated correlation of .78, which was the largest difference between observed and disattenuated correlation. In addition, scanning through Table 9.11 reveals that the pattern of convergent evidence is improved and the divergent evidence somewhat weaker using disattenuated correlations, which is indicative that test reliability may explain part of the stronger divergent evidence for mathematics and science. For example, grade 7 ARMT+ mathematics was most strongly correlated with ARMT+ science and reading in table 9.10 (divergent evidence), but the largest disattenuated grade 7 ARMT+ mathematics correlation was between mathematics scores for ARMT+ and ACT Aspire.

Table 9.11. Disattenuated Correlations between ARMT+ and ACT Aspire Scale Scores by Grade Level

		Alabama			ACT Aspire		
Grade (valid N)	**Scale Score**	**Math**	**Reading**	**Science**	**Math**	**Reading**	**Science**
Grade 3 (8,530)	Alabama Math						
	Alabama Reading	.78					
	ACT Aspire Math	**.71**	.76				
	ACT Aspire Reading	.64	**.84**		.81		
Grade 4 (8,585)	Alabama Math						
	Alabama Reading	.77					
	ACT Aspire Math	**.77**	.72				
	ACT Aspire Reading	.70	**.88**		.75		
Grade 5 (6,543)	Alabama Math						
	Alabama Reading	.80					
	Alabama Science	.76	.76				
	ACT Aspire Math	**.78**	.78	.63			
	ACT Aspire Reading	.74	**.87**	.72	.76		
	ACT Aspire Science	.79	.83	**.73**	.82	.89	
Grade 6 (5,804)	Alabama Math						
	Alabama Reading	.80					
	ACT Aspire Math	**.75**	.64				
	ACT Aspire Reading	.73	**.85**		.68		
Grade 7 (4,345)	Alabama Math						
	Alabama Reading	.79					
	Alabama Science	.80	.80				
	ACT Aspire Math	**.83**	.70	.70			
	ACT Aspire Reading	.77	**.84**	.79	.79		
	ACT Aspire Science	.80	.79	**.76**	.83	.89	
Grade 8 (3,876)	Alabama Math						
	Alabama Reading	.81					
	ACT Aspire Math	**.91**	.78				
	ACT Aspire Reading	.72	**.85**		.77		

To summarize, we see relatively strong positive relationships between ACT Aspire and ARMT+ scores in the same subject, particularly if we account for the reliabilities of scores. In addition, we see stronger convergent evidence that ARMT+ and ACT Aspire are measuring similar traits in reading. In mathematics and science, we see divergent evidence in grades 3–7 indicative of method effects. It was somewhat surprising to see persistent divergent evidence in mathematics and science. The results are consistent with an interpretation of mathematics and science subjects as being particularly distinct from one another across the ARMT+ and ACT Aspire tests, but the reliabilities of some tests likely exaggerated divergent evidence. Test design issues could also explain the divergent evidence observed in science. ARMT+ science only included selected-response items, whereas ACT Aspire also included constructed-response items. However, mathematics and reading tests for ARMT+ and ACT Aspire contained selected-response items and constructed-response items, so this explanation may not apply to the divergent evidence observed for mathematics tests.

From a validity perspective, these results support the argument that ACT Aspire and ARMT+ measure academic achievement in mathematics, reading, and science. However, these results also support the argument that ARMT+ and ACT Aspire are not the same and likely measure different aspects within the broad domain of academic achievement in each subject area.

Summary

This chapter provided some evidence regarding the validation of ACT Aspire as testing academic achievement in English, mathematics, reading, and science which is one component of validity evidence for ACT Aspire score interpretations. Much work on the validation of ACT Aspire remains. This is especially true for a new testing program like ACT Aspire where the body of evidence for the interpretations of ACT Aspire scores for particular uses must be established.

To more clearly articulate the interpretative and validity arguments for the two primary and three secondary interpretations of ACT Aspire scores and their associated uses, existing evidence is being summarized and reported and additional evidence will be collected for validation purposes. This evidence is not limited to a chapter in technical documentation titled “Validity”; it touches or relies on each chapter of this document and more. As stated by The Standards: “a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses” (AERA, APA, and NCME 2014, p. 21). These strands of evidence come from multiple sources and may involve theoretical, logical and/or empirical evidence (Kane 2006).

This chapter presented and summarized two studies that provided evidence regarding the relationships of ACT Aspire scores with scores on other assessments. However, several types of evidence need to be addressed directly to support further validation of ACT Aspire scores, including evidence based on test content, evidence based on response processes, evidence based on internal test structure, evidence based on relations to other variables (specifically, test-criterion relationships), and evidence considering the consequences of testing. Evidence will continue to be collected to support the interpretations of ACT Aspire scores.

CHAPTER 10

ACT Aspire Mode Comparability Study

Introduction

A comparability study was undertaken to investigate ACT Aspire performance across online and paper modes. Most test forms included in this study were not identical across modes. This led us to assume that scores from paper and online forms were *not* interchangeable without first statistically linking test forms across modes.

Two primary purposes of the comparability study included (1) determining whether mode of administration affected student performance on common (i.e., identical) items across modes prior to scaling and (2) investigating the comparability of ACT Aspire scale scores, which were obtained after linking paper and online forms. The overarching question to address with the comparability study was, Is the validity of the interpretations of test scores similar across paper and online test forms? In this chapter we describe comparisons of raw number-of-points scores (on identical items) and scale scores across mode. These two types of scores are likely of most interest to test users concerned about comparability across mode. Other forthcoming documentation will provide a more complete description of the comparability study.

Studying the effects of mode on performance of collections of identical items gives us an idea of direction and degree of differences due to mode. This, in turn, can help us determine whether performance on collections of identical items across mode could be considered interchangeable without statistical linking across mode to moderate potential mode effects.

Investigating the comparability of scale scores after linking forms across mode, where items on forms are not 100% the same but may have most of the same items, provides us with evidence regarding (a) the effectiveness of the linking and (b) the

apparent interchangeability of scale scores across forms administered in different modes. When a testing program maintains test forms administered in different modes and scores are to be used interchangeably across forms, it is incumbent on those maintaining the test to show that scale scores are indeed comparable across forms (e.g., see standard 5.17 of the *Standards for Educational and Psychological Testing*; AERA, APA, and NCME 2014, 106).

Method and Results

Test materials included one online and one paper ACT Aspire English, Mathematics, Reading, Science, and Writing test form for grades 3–10. At least 80% of test items were identical across mode and consisted of multiple-choice and constructed-response item types.[16] However, most online test forms also contained a small number of technology-enhanced and fill-in-the-blank items that were not amenable to paper-based presentation. Paper forms contained analogous multiple-choice items covering the same content, but these items were not considered identical across forms due to item-type differences.[17] Table 10.1 shows the percentages of items considered different across mode. We handled these nonidentical items differently depending on the purpose of our analysis. For studying mode effects on raw number-of-points scores, we excluded such items and focused on the common (identical) items to gauge mode effects. For studying the comparability of scale scores across forms (by mode) we included the nonidentical items because scale scores are obtained using all items on a test form.

Participants in this study included students in grades 3 through 10 from a sample representing 13 states, 72 districts, and 108 schools. Table 10.2 identifies sample sizes for each subject by grade and mode. Sample sizes ranged from 645 (online Grade 8 Writing) to 1,539 (paper Grade 3 Mathematics).

Table 10.1. Percentage of Items Different* Across Online and Paper Forms

	Grade						
Subject	**3**	**4**	**5**	**6**	**7**	**8**	**EHS**
English	16%	20%	16%	9%	6%	—	—
Math	20%	16%	12%	18%	15%	8%	5%
Reading	8%	8%	13%	8%	13%	—	—
Science	11%	14%	14%	13%	13%	9%	13%
Writing	—	—	—	—	—	—	—

Note: — = all items identical across modes (writing was based on a single prompt).

* Technology enhanced items and fill-in-the blank items were included in online forms, with selected response analogs on paper forms.

[16] English forms did not contain constructed response items.

[17] A separate study will explore nonoverlapping items across paper and online modes.

Table 10.2. Sample Sizes by Grade and Subject

	Grade													
	3		4		5		6		7		8		EHS	
Subject	**O**	**P**	**O**	**P**	**O**	**P**	**O**	**P**	**O**	**P**	**O**	**P**	**O**	**P**
English	1,359	1,405	1,480	1,489	1,349	1,346	1,129	1,166	1,019	1,033	866	887	1,176	1,209
Mathematics	1,488	1,539	1,534	1,532	1,450	1,427	1,156	1,194	974	1,010	896	927	1,075	1,098
Reading	1,441	1,513	1,399	1,401	1,299	1,294	1,004	1,038	1,010	1,037	800	831	857	919
Science	1,376	1,405	1,435	1,444	1,378	1,379	1,047	1,040	1,028	1,042	913	934	959	1,023
Writing	1,189	1,268	1,260	1,280	1,101	1,102	840	885	825	836	645	658	698	767

Note: O = online form, P = paper form.

Students were recruited to complete an ACT Aspire test in one or more subjects. Students at each grade within a school were randomly assigned to take either the paper or the online version of the test. This design is called *random equivalent groups* and ensured that recruited students had an equal chance of being assigned the online or paper test within a school. If students testing in each mode are equivalent, we can attribute observed differences in performance to differences in testing mode, not to differences in groups of students (or a combination of group and performance differences).

To help ensure our analysis included adequately balanced samples of students testing in each mode, schools were included in the analysis sample only if the ratio of students testing in one mode compared to the other within the school was less than two, implying that fewer than twice as many students tested in one mode versus the other. This data cleaning rule led to excluding fewer than 10% of students from analysis for most grades and subjects.[18]

Mode Effects for Raw Number-of-Points Scores on Common Items

We investigated student raw score performance by summing scores across items that were the same across paper and online forms (i.e., excluding those items mentioned earlier that differed in item type across mode). English and writing showed more consistent evidence of mode effects across grades compared to mathematics, reading, and science. But for some subjects and grades, raw scores appeared to differ across mode, and for others they did not. It is apparent from these results that it was not safe to assume raw scores and raw score distributions on identical items across modes were comparable.

Tables 10.3–10.7 summarize the aggregated raw score statistics for the common items across mode, including score moments (mean, standard deviation, skewness, and kurtosis), effect sizes (standardized differences in mean scores), and statistical significance tests from tests of equivalence (for details on this test, see Rogers,

[18] Up to 22% of students (five schools) were excluded from analysis. Grade 7 Science and Grade 8 Writing had more than 20% of students excluded. Most excluded schools only tested in one mode.

Howard, and Vessey 1993; Serlin and Lapsley 1985; see also the Kolmogorov-Smirnov test of differences in cumulative distributions by subject).

Table 10.3. English Common-Item Raw Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	21	21	20	20	21	21	32	32	33	33	35	35	50	50
Sample size	1,359	1,405	1,480	1,489	1,349	1,346	1,129	1,166	1,019	1,033	866	887	1,176	1,209
Raw score														
Mean	11.65	10.07	11.44	11.06	11.21	10.61	16.52	14.84	17.44	16.41	21.91	22.06	25.75	25.54
SD	4.14	4.05	3.50	3.77	4.15	4.04	5.32	5.32	5.92	5.85	5.97	6.52	9.96	9.81
Minimum	1	1	1	1	0	0	3	0	4	2	5	3	4	3
Maximum	21	20	20	20	21	20	32	30	32	33	34	34	49	50
Skewness	0.00	0.36	-0.09	-0.11	0.11	0.09	0.14	0.30	0.02	0.24	-0.32	-0.43	0.20	0.16
Kurtosis	-0.68	-0.64	-0.65	-0.57	-0.71	-0.67	-0.34	-0.38	-0.62	-0.49	-0.51	-0.54	-0.82	-0.75
Reliability	.79	.78	.75	.78	.79	.78	.80	.80	.81	.81	.83	.86	.91	.90
Effect size	0.39		0.11		0.15		0.32		0.18		-0.02		0.02	
Stat. test														
Kolmogorov-Smirnov test	4.68*		1.45*		1.66*		3.47*		1.90*		1.12		0.56	
Test of equivalence for mean difference† *t*-lower/*t*-upper	16.58/3.73		10.35/-4.62*		10.14/-2.54*		12.08/3.08		7.84/0.14		2.84/-3.85*		2.99/-1.96*	
(df)	(2,762)		(2,967)		(2,693)		(2,293)		(2,050)		(1,751)		(2,383)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.4. Mathematics Common-Item Raw Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	20	20	21	21	22	22	28	28	29	29	35	35	36	36
Sample size	1,488	1,539	1,534	1,532	1,450	1,427	1,156	1,194	974	1,010	896	927	1,075	1,098
Raw score														
Mean	12.42	12.61	9.53	9.48	10.00	9.99	12.27	12.01	13.31	13.91	18.96	19.75	17.60	18.14
SD	4.68	4.77	3.46	3.39	3.49	3.34	4.13	4.11	5.51	5.57	7.16	7.18	8.21	8.15
Minimum	0	1	1	0	1	0	2	1	2	1	1	1	3	1
Maximum	26	27	25	27	26	23	31	29	34	32	42	39	44	45
Skewness	0.17	0.16	0.60	0.62	0.58	0.32	0.53	0.36	0.61	0.55	0.25	0.03	0.48	0.45
Kurtosis	-0.29	-0.41	1.00	1.12	0.69	0.24	0.51	0.25	0.20	-0.03	-0.37	-0.55	-0.47	-0.46
Reliability	.76	.75	.64	.62	.67	.63	.74	.73	.82	.83	.87	.87	.88	.87
Effect size	-0.04		0.02		0.00		0.06		-0.11		-0.11		-0.07	
Stat. test														
Kolmogorov-Smirnov test	1.05		0.43		0.96		0.84		1.25		1.47*		0.73	
Test of equivalence for mean difference† *t*-lower/*t*-upper	4.74/-6.91*		8.54/-7.63*		7.92/-7.78*		7.42/-4.34*		1.62/-6.41		0.63/-5.32		1.32/-4.38	
(df)	(3,025)		(3,064)		(2,875)		(2,348)		(1,982)		(1,821)		(2,171)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.5. Reading Common-Item Raw Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	22	22	22	22	21	21	22	22	21	21	24	24	24	24
Sample size	1,441	1,513	1,399	1,401	1,299	1,294	1,004	1,038	1,010	1,037	800	831	857	919
Raw score														
Mean	12.93	13.99	13.30	14.32	12.94	12.82	14.56	14.69	11.96	12.31	16.39	15.96	15.17	15.20
SD	5.75	5.68	5.66	5.69	5.13	5.03	5.44	5.15	5.20	5.31	6.25	6.41	6.87	6.99
Minimum	1	1	1	1	1	1	2	2	1	0	1	1	1	1
Maximum	26	26	26	26	25	24	26	26	25	24	29	30	30	31
Skewness	0.09	-0.15	0.10	-0.14	0.05	-0.08	-0.11	-0.21	0.15	0.04	-0.04	-0.16	0.01	-0.02
Kurtosis	-1.01	-0.88	-0.91	-0.87	-0.82	-0.81	-0.91	-0.82	-0.81	-0.94	-0.88	-0.86	-1.07	-1.10
Reliability	.85	.84	.83	.83	.83	.83	.83	.81	.80	.81	.83	.84	.87	.88
Effect size	-0.19		-0.18		0.02		-0.03		-0.07		0.07		0.00	
Stat. test														
Kolmogorov-Smirnov test	2.76*		2.48*		0.52		0.68		0.96		0.73		0.47	
Test of equivalence for mean difference† *t*-lower/*t*-upper	-0.31/-9.81		-0.09/-9.42		5.59/-4.44*		3.68/-4.85*		2.79/-5.81*		4.57/-1.81*		2.94/-3.14*	
(df)	(2,952)		(2,798)		(2,591)		(2,040)		(2,045)		(1,629)		(1,774)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.6. Science Common-Item Raw Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	25	25	24	24	24	24	28	28	28	28	29	29	28	28
Sample size	1,376	1,405	1,435	1,444	1,378	1,379	1,047	1,040	1,028	1,042	913	934	959	1,023
Raw score														
Mean	12.43	12.69	13.94	14.15	15.17	14.97	17.25	17.04	14.95	15.18	16.80	16.93	13.18	14.04
SD	7.07	6.77	5.63	5.74	6.28	6.25	7.30	7.30	7.59	7.34	7.11	7.28	6.80	7.20
Minimum	0	1	1	1	1	1	2	1	1	2	2	2	2	1
Maximum	32	32	30	30	32	29	35	35	35	33	36	36	35	36
Skewness	0.52	0.52	0.27	0.23	0.02	0.02	0.21	0.09	0.42	0.35	0.19	0.15	0.76	0.61
Kurtosis	-0.60	-0.51	-0.54	-0.57	-0.80	-0.84	-0.79	-0.86	-0.80	-0.88	-0.78	-0.71	-0.21	-0.33
Reliability	.88	.88	.84	.84	.85	.85	.88	.88	.89	.89	.87	.87	.87	.87
Effect size	-0.04		-0.04		0.03		0.03		-0.03		-0.02		-0.12	
Stat. test														
Kolmogorov-Smirnov test	1.22		0.71		0.54		0.59		0.89		0.59		1.69*	
Test of equivalence for mean difference† *t*-lower/*t*-upper	2.81/-4.81*		3.73/-5.71*		5.02/-3.36*		3.77/-2.48*		2.34/-3.75*		2.60/-3.37*		0.46/-5.89	
(df)	(2,779)		(2,877)		(2,755)		(2,085)		(2,068)		(1,845)		(1,980)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.7. Writing Common-Item Raw Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sample size	1,189	1,268	1,260	1,280	1,101	1,102	840	885	825	836	645	658	698	767
Raw score														
Mean	11.1	11.8	11.18	10.72	11.44	12.21	13.55	14.48	12.38	13.29	12.18	11.94	12.38	12.27
SD	3.53	3.01	3.56	3.13	3.74	3.27	3.95	3.44	3.87	3.33	3.5	3.18	3.65	3.57
Minimum	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Maximum	20	20	20	20	20	20	23	24	24	24	23	24	24	24
Skewness	0.21	0.03	-0.07	0.23	0.08	-0.13	-0.14	-0.37	0.27	0.17	-0.20	-0.01	0.02	0.04
Kurtosis	-0.37	-0.20	-0.52	-0.33	-0.28	-0.16	-0.32	-0.03	0.09	0.21	-0.20	0.31	0.12	0.10
Effect size	-0.21		0.14		-0.22		-0.25		-0.25		0.07		0.03	
Stat. test														
Kolmogorov-Smirnov test	3.27*		2.40*		2.65*		3.12*		2.64*		1.59*		0.86	
Test of equivalence for mean difference† *t*-lower/*t*-upper	2.29/-12.85*		10.92/-4.11*		1.49/-11.87		0.39/-10.84		0.49/-10.8		6.67/-4.14*		5.87/-4.72*	
(df)	(2,455)		(2,538)		(2,201)		(1,723)		(1,659)		(1,301)		(1,463)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

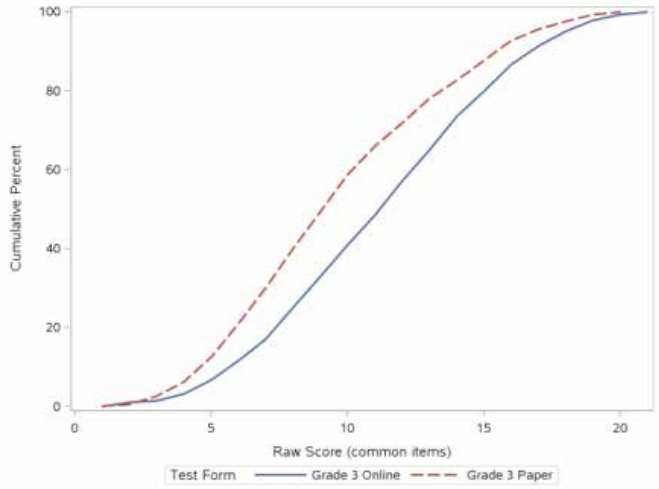
† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Figure 10.1 displays the cumulative percent of students at each raw score by mode for each subject and grade. The solid curve represents the online form and the dashed curve represents the paper form. When one of the plotted cumulative percent curves is to the left of the other, it indicates that this group scored lower relative to the group to the right (or, conversely, the group to the right scored higher). If curves cross or if the relative position of curves varies across raw scores, it indicates that the cumulative percent of students between groups varies across scores. The distance between the two curves indicates the magnitude of difference between modes (difference in cumulative percent in the vertical direction, difference in raw score in the horizontal direction). For example, for Grade 3 English, the online cumulative percent curve is to the right of the paper curve, which implies that online students scored higher. Reading up from a raw score point of 10, approximately 40% of online students scored 10 or lower, whereas approximately 60% of paper students scored 10 or lower. By subtracting these percentages from 100, one could say that approximately 60% of online students scored above a 10, but about 40% of students testing on paper scored above a 10. A statistical test of the differences in cumulative distributions is provided by the Kolmogorov-Smirnov test listed in table 10.3.

Figure 10.1. Plots of cumulative percent of students for common item raw scores across mode

English

Grade 3



Grade 4

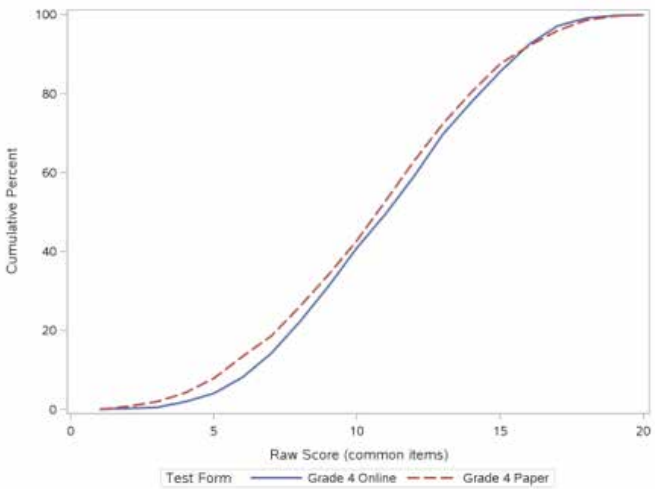
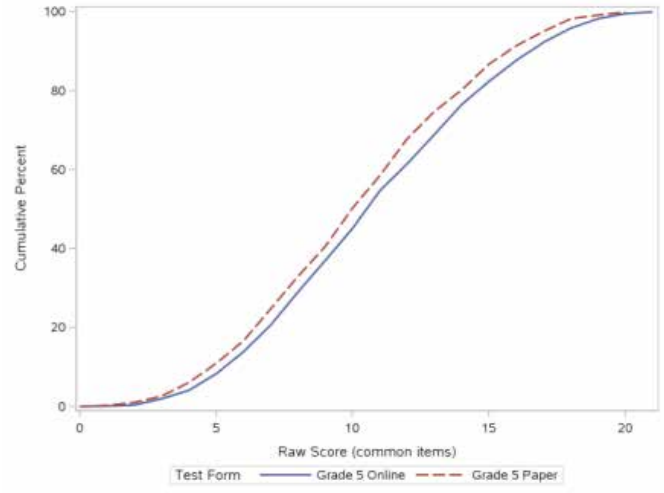
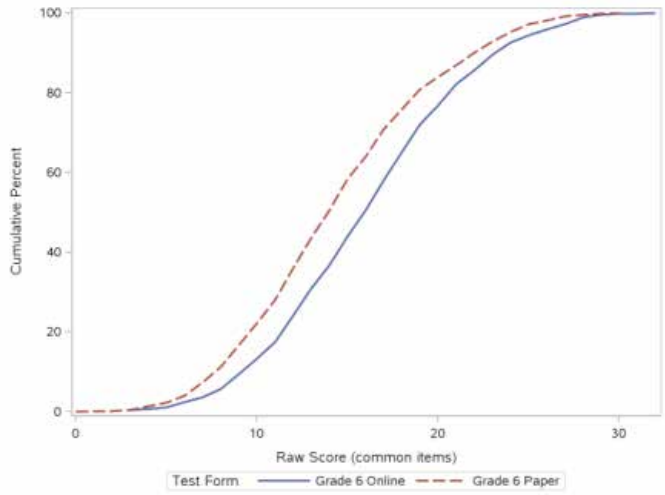


Figure 10.1. (continued)

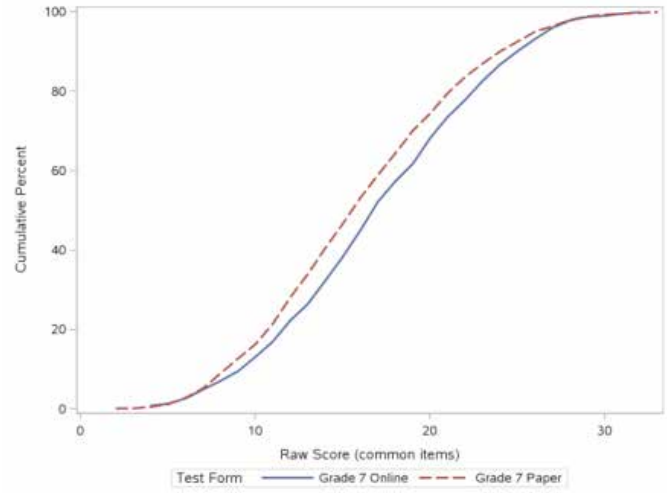
Grade 5



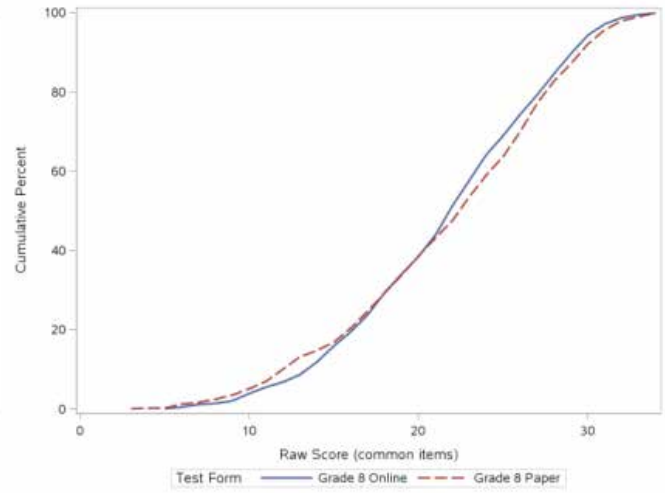
Grade 6



Grade 7



Grade 8



Early High School

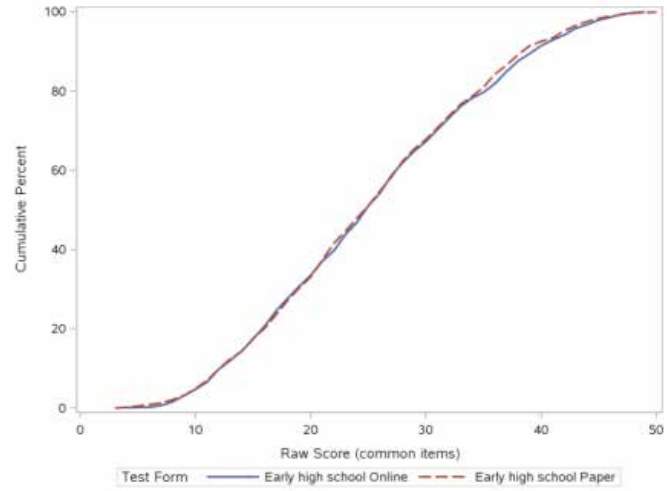
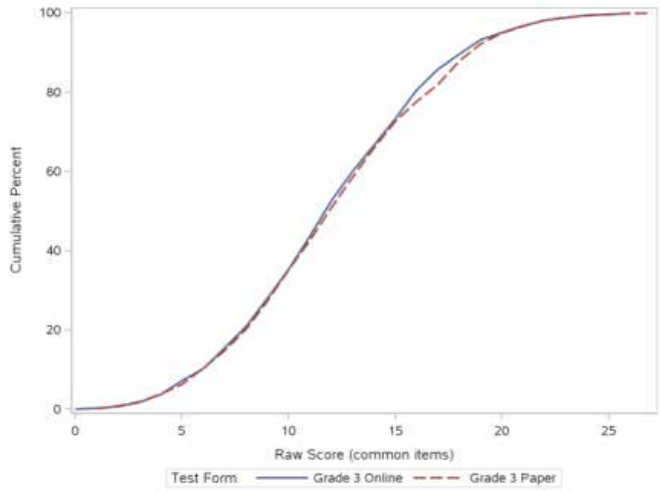


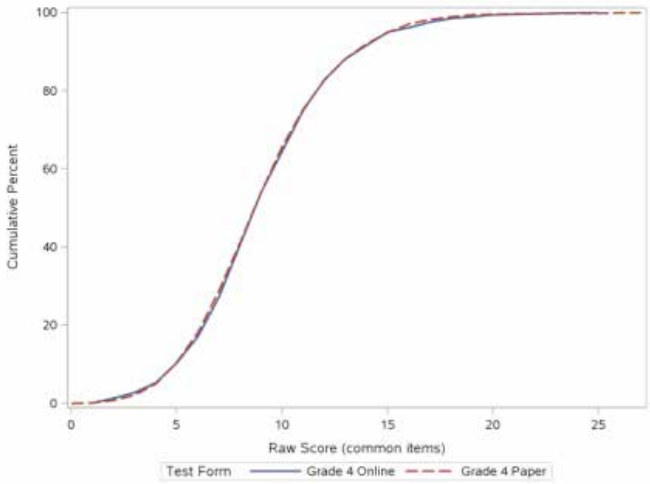
Figure 10.1. (continued)

Mathematics

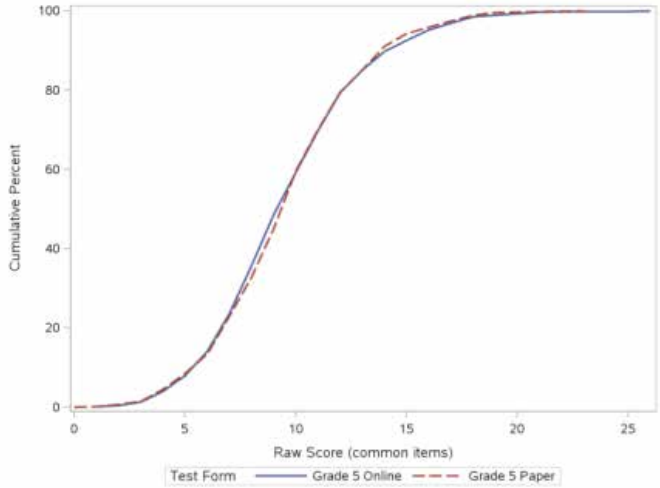
Grade 3



Grade 4



Grade 5



Grade 6

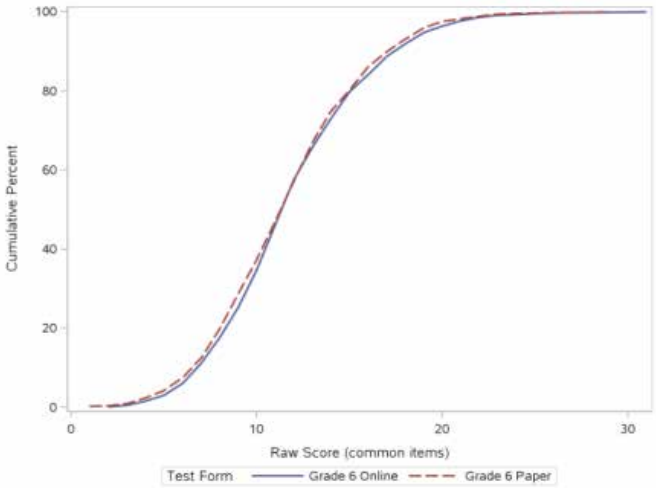
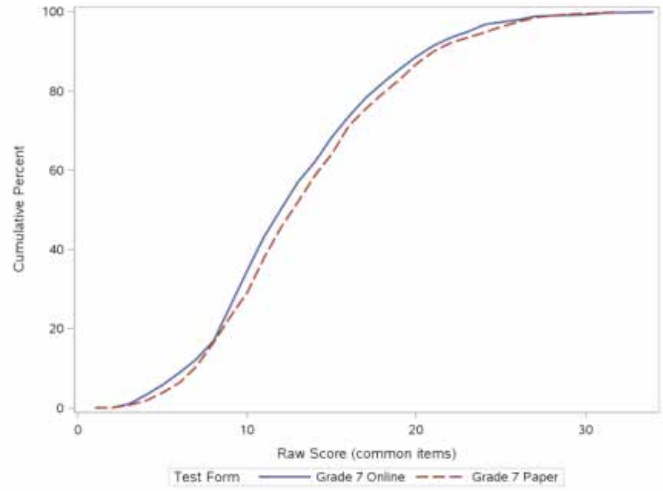
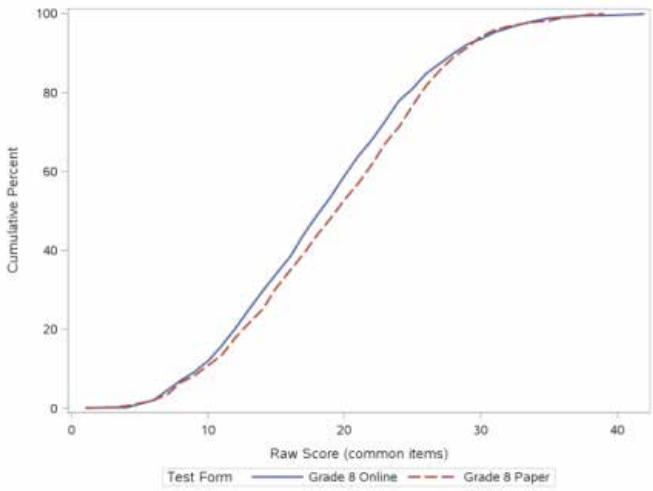


Figure 10.1. (continued)

Grade 7



Grade 8



Early High School

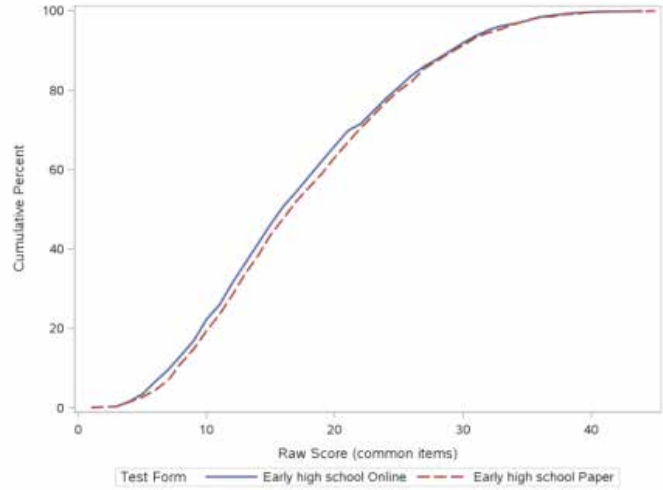
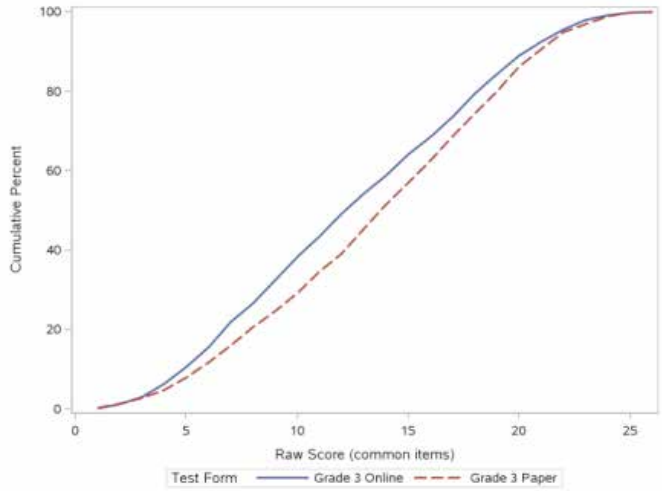


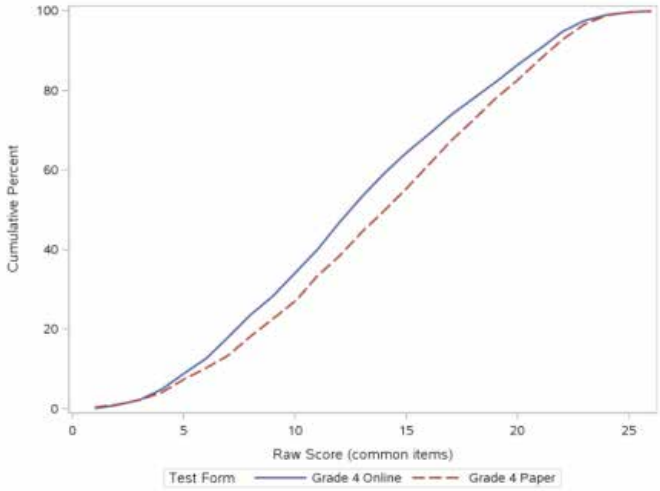
Figure 10.1. (continued)

Reading

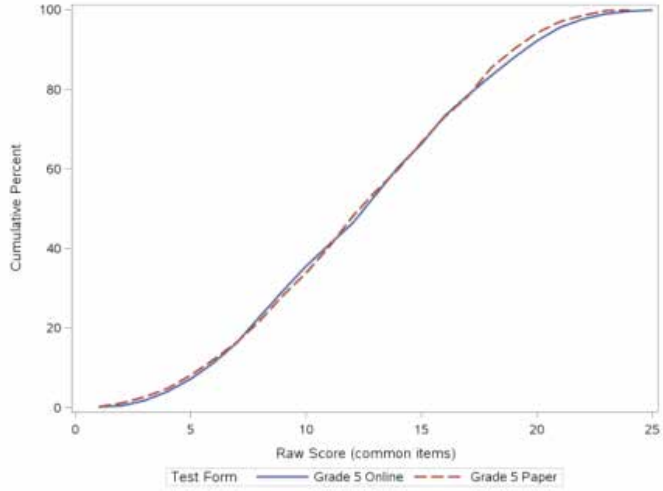
Grade 3



Grade 4



Grade 5



Grade 6

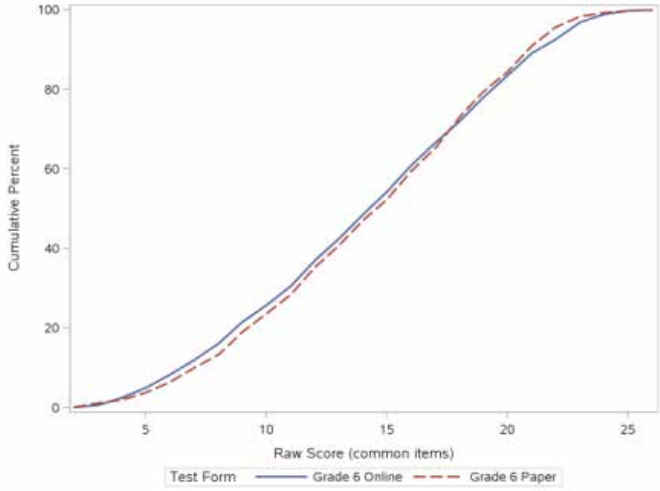
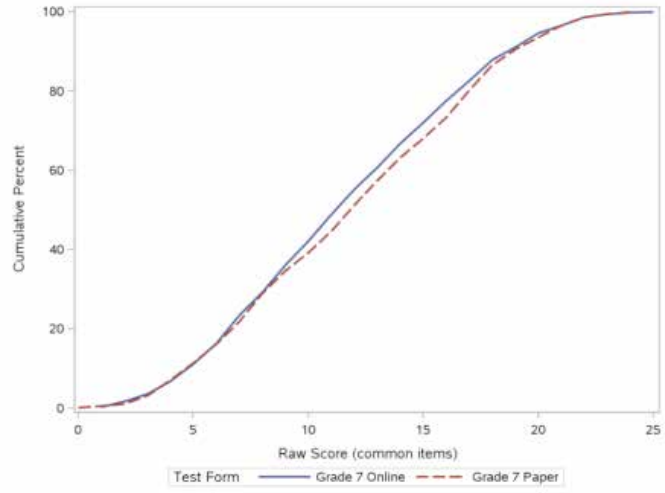
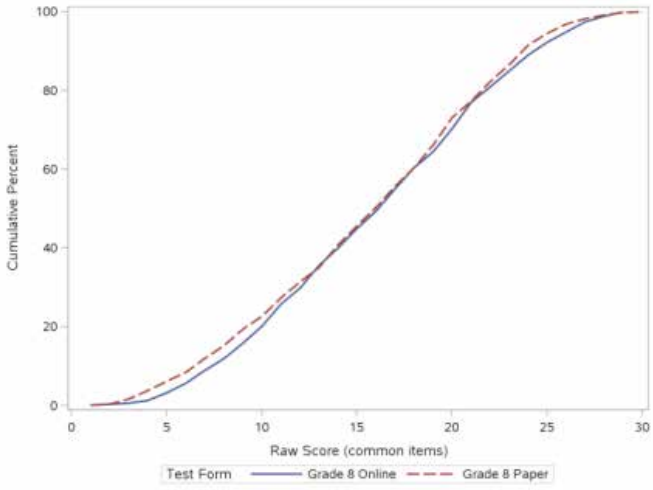


Figure 10.1. (continued)

Grade 7



Grade 8



Early High School

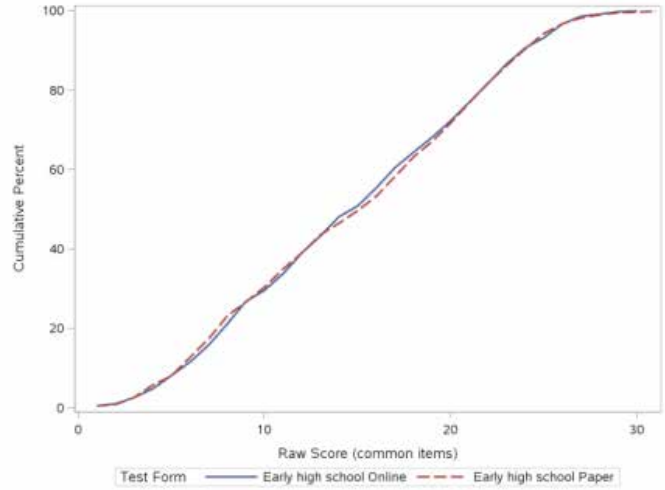
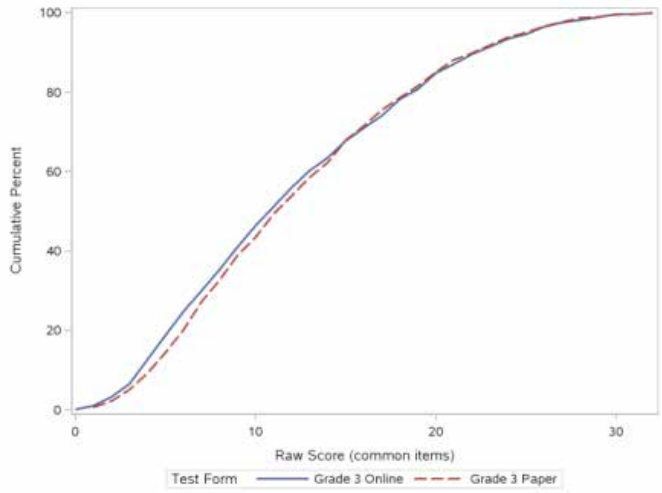


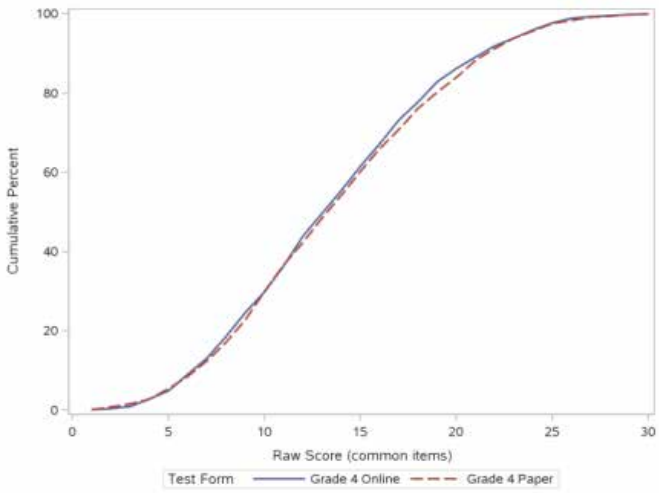
Figure 10.1. (continued)

Science

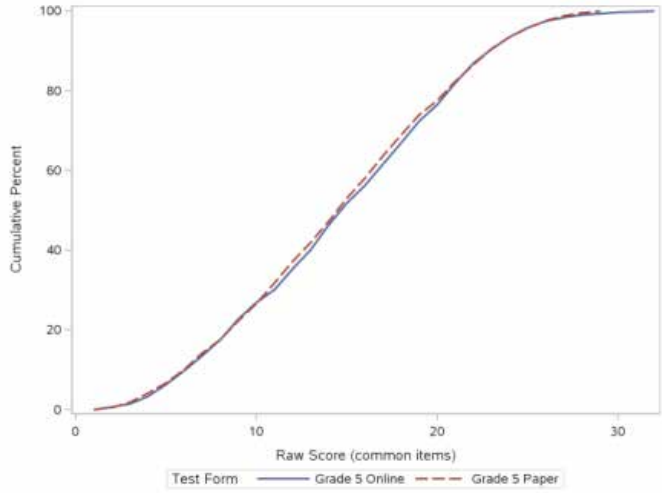
Grade 3



Grade 4



Grade 5



Grade 6

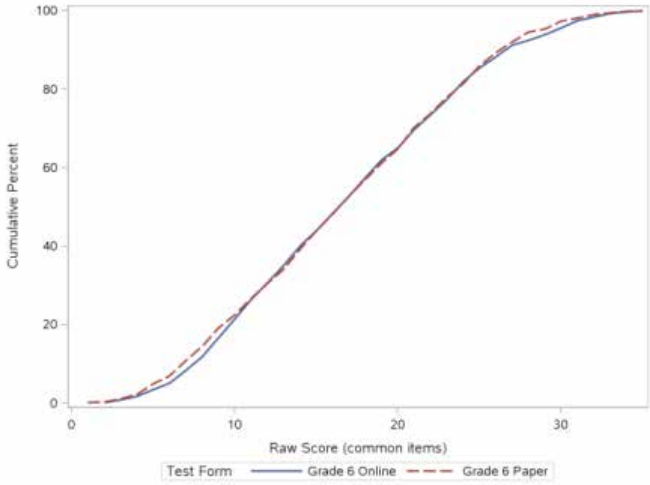
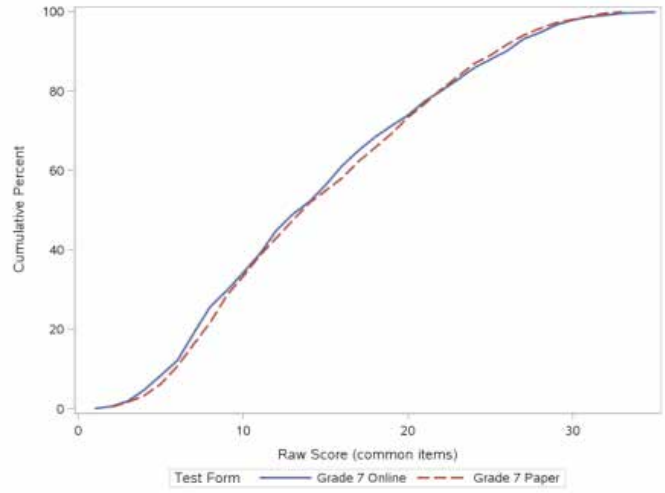
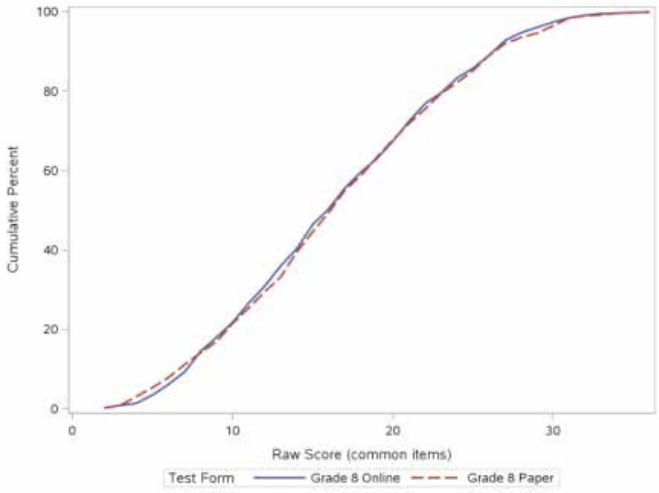


Figure 10.1. (continued)

Grade 7



Grade 8



Early High School

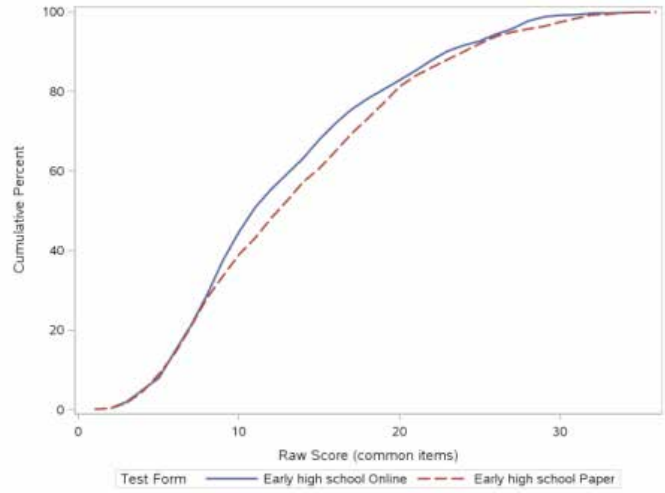
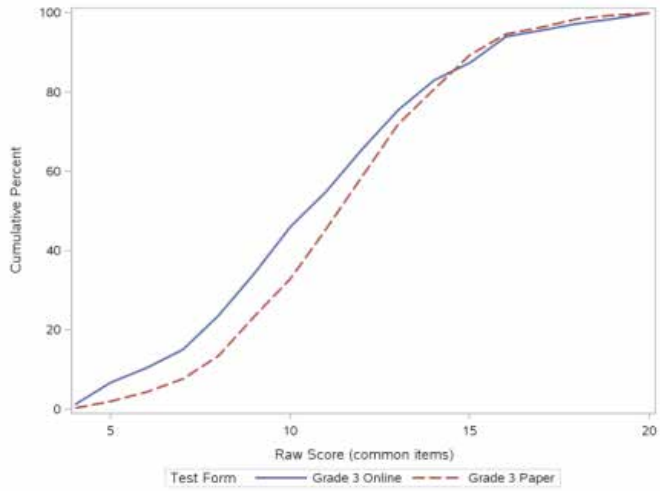


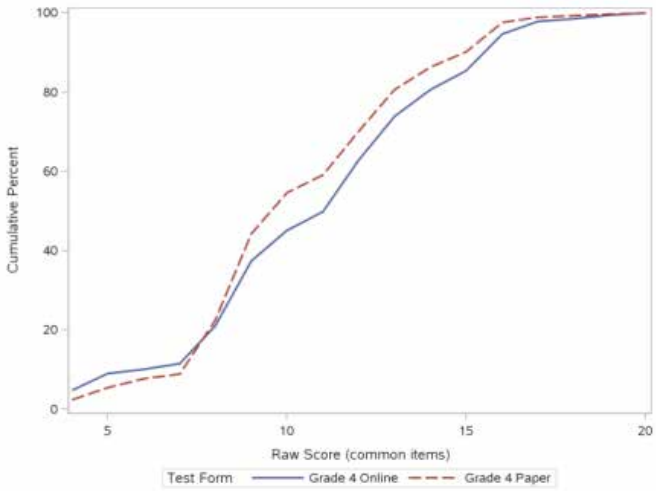
Figure 10.1. (continued)

Writing

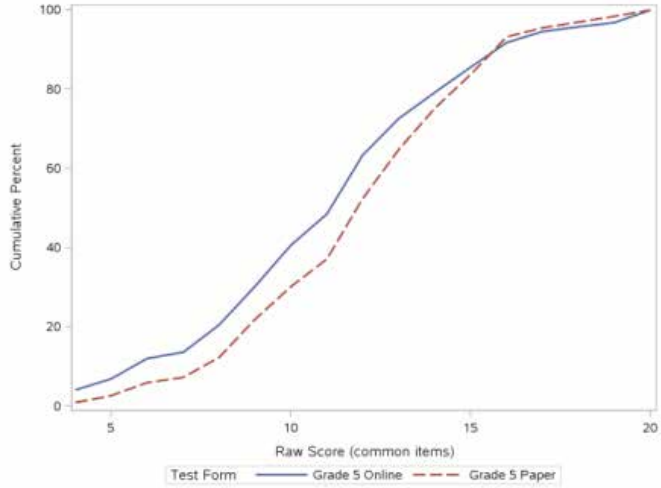
Grade 3



Grade 4



Grade 5



Grade 6

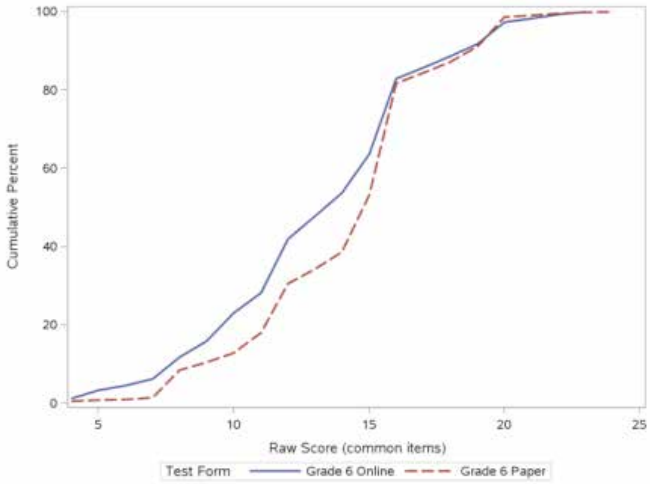
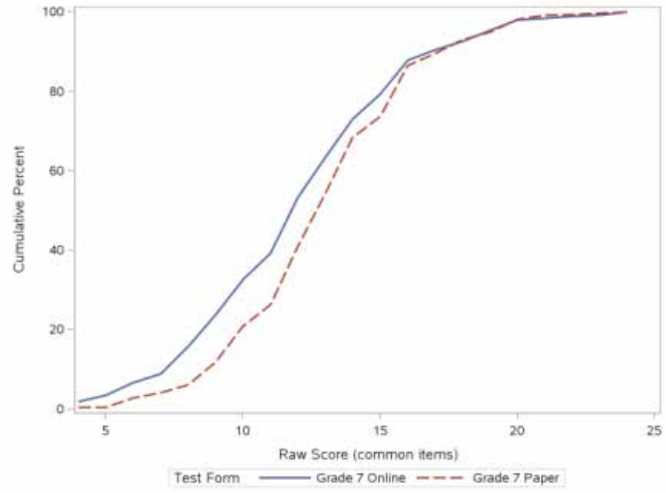
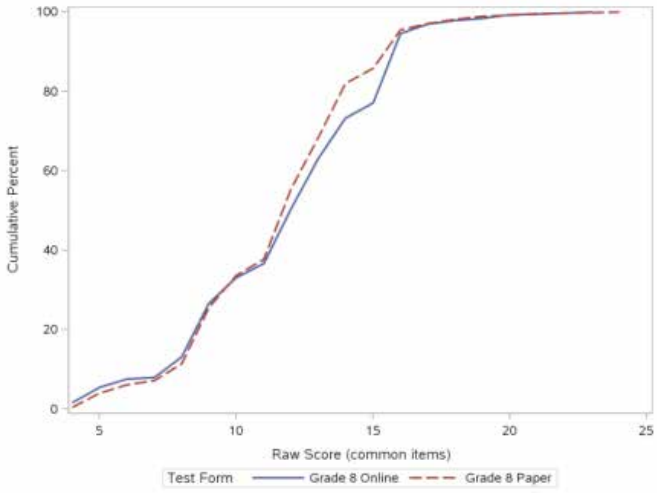


Figure 10.1. (continued)

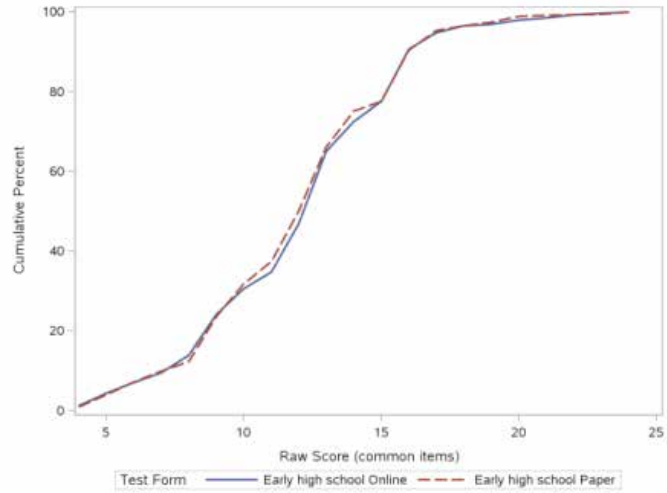
Grade 7



Grade 8



Early High School



English raw scores and distributions showed evidence of differences across modes in grades 3, 5, 6, and 7, with means differing by more than one point, effect sizes between .15 and .39,[19] statistically significant differences in score distributions, and nonstatistically significant similarities in mean scores. Although not consistent across statistics, grade 4 did show some evidence of mode effects with a statistically significant Kolmogorov-Smirnov test indicating that score distributions differed (see also figure 10.1 for a plot of the score distributions). Online scored higher than paper for grades where differences were evident.

Mathematics raw scores and distributions showed evidence of differences across modes in grade 7, 8, and EHS, with the paper group scoring higher than the online group. The test of equivalence indicated that mean scores were not statistically similar. The score distributions differed statistically at grade 8 (see figure 10.1, which plots the difference). However, the magnitudes of differences across mode were relatively small; means differed by less than one point and effect sizes were within ±.15.

Reading raw scores and distributions showed evidence of differences across modes in grades 3 and 4, with the paper group scoring higher than the online group. Means differed by more than one point, effect sizes were −.19 and −.18, differences in score distributions were statistically significant, and similarities in mean scores were not statistically significant.

Science raw scores and distributions did not appear to differ across mode for all but the EHS forms, where distributions were statistically different and means were not statistically similar. However, the difference was not large; means favored paper by less than one point and the effect size was −.12.

For writing raw scores and distributions, results showed differences across modes for all forms except EHS. Forms showed statistically significant differences in distributions in grades 3–8. Tests of equivalence did not show evidence of similarity of means for grades 3, 4, 5, 6, or 7. However, means differed by less than one point for all but grade 7. Effect sizes were within ±.15 for grades 4, 8 and EHS and were between −.20 and −.30 for grades 3, 4, 5, and 7. Paper scored higher than online in grades 3, 5, 6, and 7 and online scored higher than paper in grades 4 and 8.

In many cases, the different methods of checking mode effects for raw number-of-points scores on common items led to slightly different grades in each subject showing statistically significant mode effects. While typically not large, there were statistically significant differences across modes observed in each subject area.

[19] We judged effect sizes within ±.15 to be small to negligible.

Our interpretation of the comparisons of raw number-of-points scores based on identical items across modes is that mode effects observed for these ACT Aspire forms should not be ignored. Score differences were not always observed, and when they were observed they were generally not large, but we would argue that it is not reasonable to assume no mode effects across collections of identical items. If two ACT Aspire forms did contain 100% overlapping items, we would recommend additional statistical linking to adjust for mode effects. However, as mentioned earlier, most ACT Aspire online and paper forms are not 100% identical, so regardless of mode effects, best practices would involve linking to place forms on the same scale (Kingston 2009).

Comparisons of Scale Scores across Mode

We investigated the comparability of ACT Aspire scale scores, which were obtained after linking paper and online forms. Paper forms were linked to online forms using random equivalent groups equipercentile linking (Kolen and Brennan 2014). Equipercentile methodology has been used extensively with other ACT testing programs, including the operational equating of ACT Aspire (see chapter 11). The primary purpose of this particular linking was to statistically adjust for differences across forms due to item type and mode.

Tables 10.8–10.12 summarize scale score statistics, including score moments (mean, standard deviation, skewness, and kurtosis), effect sizes, statistical significance tests from tests of equivalence, and the Kolmogorov-Smirnov test of difference in cumulative distributions for each subject. Figure 10.2 displays the cumulative percent of students at each scale score by mode for each subject and grade. Figure 10.3 displays box plots of scale scores by mode for each subject and grade.

Table 10.8. English Scale Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	25	25	25	25	25	25	35	35	35	35	35	35	50	50
Sample size	1,359	1,405	1,480	1,489	1,349	1,346	1,129	1,166	1,019	1,033	866	887	1,176	1,209
Scale score														
Mean	416.32	416.38	419.64	419.52	421.79	421.73	424.23	424.22	424.74	424.63	427.61	427.64	428.23	428.05
SD	5.87	5.86	6.14	6.13	7.15	7.05	7.73	7.80	8.93	8.95	8.74	8.79	10.75	10.66
Minimum	401	401	402	405	403	404	402	400	403	400	403	401	400	400
Maximum	435	434	438	438	442	441	447	446	448	450	448	447	454	456
Skewness	0.51	0.49	0.08	0.11	0.23	0.25	0.12	0.12	-0.08	-0.07	-0.17	-0.17	0.01	0.01
Kurtosis	-0.09	-0.05	-0.50	-0.58	-0.44	-0.35	-0.21	-0.24	-0.39	-0.34	-0.40	-0.39	-0.69	-0.60
Reliability	0.79	0.79	0.76	0.79	0.80	0.79	0.82	0.81	0.83	0.82	0.84	0.86	0.90	0.90
Effect size	-0.01		0.02		0.01		0.00		0.01		0.00		0.02	
Stat. test														
Kolmogorov-Smirnov test	1.41		1.26		1.42		1.11		1.02		0.94		0.64	
Test of equivalence for mean difference† *t*-lower/*t*-upper	-4.79/4.17*		-3.91/4.97*		-3.44/3.87*		-3.07/3.10*		-2.27/2.80*		-2.46/2.32*		-1.87/2.69*	
(df)	(2,762)		(2,967)		(2,693)		(2,293)		(2,050)		(1,751)		(2,383)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.9. Mathematics Scale Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	20	20	21	21	22	22	28	28	29	29	35	35	36	36
Sample size	1,488	1,539	1,534	1,532	1,450	1,427	1,156	1,194	974	1,010	896	927	1,075	1,098
Scale score														
Mean	412.21	412.19	415.25	414.99	417.44	417.37	419.85	419.87	419.73	420.03	422.92	422.65	424.88	425.11
SD	3.96	4.04	4.15	4.18	4.83	4.75	5.69	5.68	6.53	6.67	7.92	7.90	8.65	8.73
Minimum	400	400	403	402	404	403	404	404	404	403	401	403	408	405
Maximum	425	427	430	433	438	435	445	442	443	442	445	444	451	451
Skewness	0.06	-0.04	0.09	0.19	0.53	0.39	0.45	0.40	0.46	0.39	0.24	0.32	0.42	0.35
Kurtosis	-0.02	-0.03	0.26	0.58	0.58	0.36	0.27	0.20	0.08	0.02	-0.37	-0.41	-0.50	-0.56
Reliability	0.79	0.77	0.68	0.66	0.70	0.69	0.78	0.77	0.84	0.83	0.89	0.88	0.89	0.88
Effect size	0.00		0.06		0.02		0.00		-0.05		0.03		-0.03	
Stat. test														
Kolmogorov-Smirnov test	1.59*		2.49*		1.95*		1.50*		0.98		0.92		1.03	
Test of equivalence for mean difference† *t*-lower/*t*-upper	-6.75/6.99*		-4.94/8.36*		-5.17/6.03*		-4.33/4.20*		-4.39/2.35*		-1.97/3.43*		-3.29/2.07*	
(df)	(3,025)		(3,064)		(2,875)		(2,348)		(1,982)		(1,821)		(2,171)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.10. Reading Scale Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	22	22	22	22	21	21	22	22	21	21	24	24	24	24
Sample size	1,441	1,513	1,399	1,401	1,299	1,294	1,004	1,038	1,010	1,037	800	831	857	919
Scale score														
Mean	411.84	411.78	414.35	414.37	416.27	416.38	418.15	418.19	419.09	419.07	421.18	421.07	421.48	421.51
SD	5.28	5.22	5.56	5.62	6.29	6.10	6.68	6.65	6.93	6.86	7.49	7.55	7.87	7.89
Minimum	401	401	402	401	401	401	403	402	402	402	401	402	403	403
Maximum	427	429	431	431	434	433	436	436	438	436	437	440	440	442
Skewness	0.42	0.39	0.26	0.27	0.23	0.28	0.05	0.04	0.04	0.03	-0.05	-0.05	-0.01	0.02
Kurtosis	-0.49	-0.50	-0.59	-0.56	-0.57	-0.52	-0.68	-0.67	-0.68	-0.71	-0.80	-0.78	-0.90	-0.89
Reliability	0.85	0.84	0.84	0.84	0.84	0.83	0.84	0.82	0.83	0.83	0.85	0.85	0.88	0.88
Effect size	0.01		0.00		-0.02		-0.01		0.00		0.01		0.00	
Stat. test														
Kolmogorov-Smirnov test	1.07		1.03		0.76		1.15		0.90		1.03		0.64	
Test of equivalence for mean difference† *t*-lower/*t*-upper	-4.90/5.45*		-4.84/4.63*		-4.57/3.65*		-3.53/3.24*		-3.21/3.35*		-2.40/2.97*		-2.74/2.60*	
(df)	(2,952)		(2,798)		(2,591)		(2,040)		(2,045)		(1,629)		(1,774)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.11. Science Scale Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	25	25	24	24	24	24	28	28	28	28	29	29	28	28
Sample size	1,376	1,405	1,435	1,444	1,378	1,379	1,047	1,040	1,028	1,042	913	934	959	1,023
Scale score														
Mean	413.60	413.59	417.02	417.09	418.98	418.83	419.69	419.71	419.70	419.74	422.97	422.80	424.49	424.64
SD	5.94	5.93	6.59	6.69	6.55	6.66	7.15	7.04	8.04	8.03	7.90	7.87	8.15	8.09
Minimum	401	401	401	400	402	401	403	401	401	402	403	403	405	404
Maximum	431	432	434	435	438	435	438	439	441	438	444	444	447	449
Skewness	0.32	0.40	0.07	0.02	-0.10	-0.10	0.07	0.10	0.17	0.16	-0.06	-0.02	0.17	0.17
Kurtosis	-0.60	-0.50	-0.55	-0.77	-0.48	-0.68	-0.65	-0.55	-0.89	-0.93	-0.67	-0.70	-0.69	-0.58
Reliability	0.89	0.88	0.85	0.85	0.86	0.87	0.89	0.88	0.90	0.89	0.88	0.88	0.86	0.87
Effect size	0.00		-0.01		0.02		0.00		0.00		0.02		-0.02	
Stat. test														
Kolmogorov-Smirnov test	1.16		1.19		0.79		0.81		0.55		0.86		0.94	
Test of equivalence for mean difference† *t*-lower/*t*-upper	-4.41/4.48*		-4.32/3.77*		-3.36/4.58*		-3.29/3.15*		-2.94/2.72*		-2.27/3.18*		-3.15/2.33*	
(df)	(2,779)		(2,877)		(2,755)		(2,085)		(2,068)		(1,845)		(1,980)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Table 10.12. Writing Scale Score Summary for Online and Paper Modes

	Grade													
	3		4		5		6		7		8		EHS	
	O	P	O	P	O	P	O	P	O	P	O	P	O	P
Number of common items	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sample size	1,236	1,295	1,272	1,289	1,106	1,106	847	888	834	839	646	665	708	777
Scale score														
Mean	422.20	422.19	422.35	422.47	422.87	422.88	427.10	427.04	424.76	424.76	424.36	424.37	424.75	424.75
SD	7.06	6.91	7.12	6.87	7.48	7.45	7.90	7.61	7.74	7.73	6.99	7.03	7.31	7.29
Minimum	408	408	408	408	408	408	408	408	408	408	408	408	408	408
Maximum	440	440	440	440	440	440	446	448	448	448	446	448	448	448
Skewness	0.21	0.23	-0.07	-0.05	0.08	0.06	-0.14	-0.06	0.27	0.28	-0.20	-0.14	0.02	0.06
Kurtosis	-0.37	-0.34	-0.52	-0.40	-0.28	-0.33	-0.32	-0.32	0.09	-0.01	-0.20	-0.17	0.12	0.13
Effect size	0.00		-0.02		0.00		0.01		0.00		0.00		0.00	
Stat. test														
Kolmogorov-Smirnov test	1.70*		3.77*		1.83*		3.76*		1.88*		1.57*		1.22	
Test of equivalence for mean difference† *t*-lower/*t*-upper	-3.50/3.60*		-4.02/3.18*		-3.17/3.12*		-2.51/2.84*		-2.65/2.62*		-2.62/2.53*		-2.62/2.62*	
(df)	(2,455)		(2,538)		(2,201)		(1,723)		(1,659)		(1,301)		(1,463)	

Note: O = online form, P = paper form.

* Statistically significant at .05.

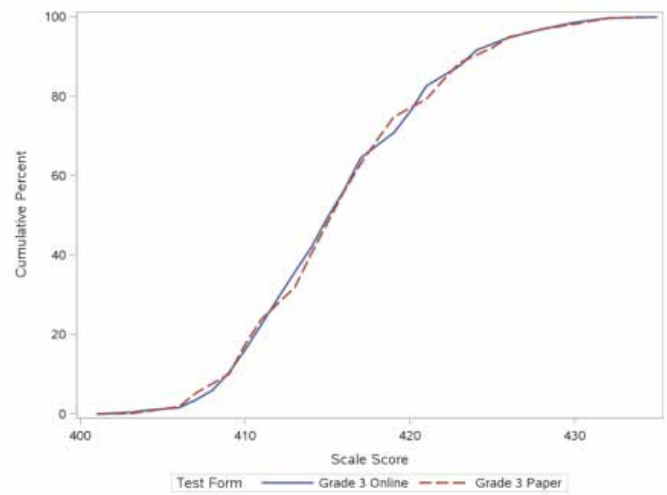
† Note that statistical significance indicates *no difference* in means across mode. This test is comprised of two one sided *t*-tests. *t-lower* is the *t*-statistic for the difference in means greater than -1 and *t-upper* is the *t*-statistic for the difference in means less than 1. Together, these two statistics test the null hypothesis that the difference in means is less than -1 or greater than 1. If both *t*-tests are statistically significant, then the difference between online and paper means is between -1 and 1.

Compared to tables 10.3–10.7, which contained raw scores on common items across mode, we see that differences in scale scores are small and generally not statistically significant. For example, if we compare the Grade 3 English raw score cumulative distributions in figure 10.1 to the Grade 3 English scale score cumulative distributions in figure 10.2, we see that the cumulative score differences are small or negligible for the scale scores. While the score differences across mode were eliminated for most of the statistics in tables 10.3–10.7, there were some cases where the Kolmogorov-Smirnov test indicated statistical differences in the cumulative distributions between modes in Mathematics and Writing. However, these differences appeared to be small, as illustrated by figure 10.2, and effect sizes were near zero. The box plots in figure 10.3 further illustrate scale score comparability across modes. The box plots are not identical across modes for each grade, but they do appear similar in most cases, which is indicative of similar score distributions across modes.

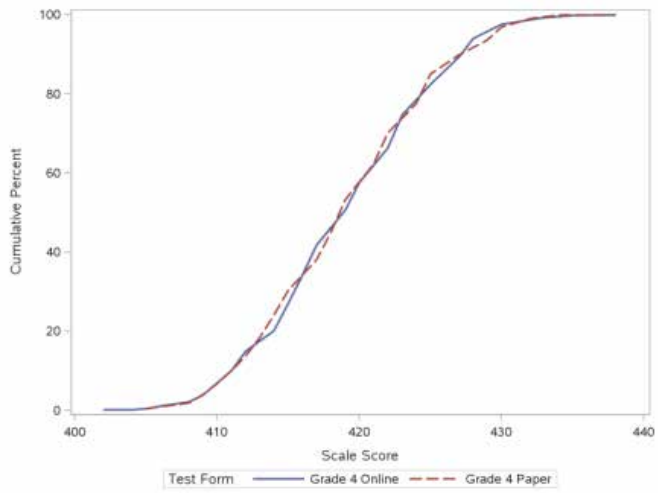
Figure 10.2. Plots of cumulative percent of students for scale scores across mode

English

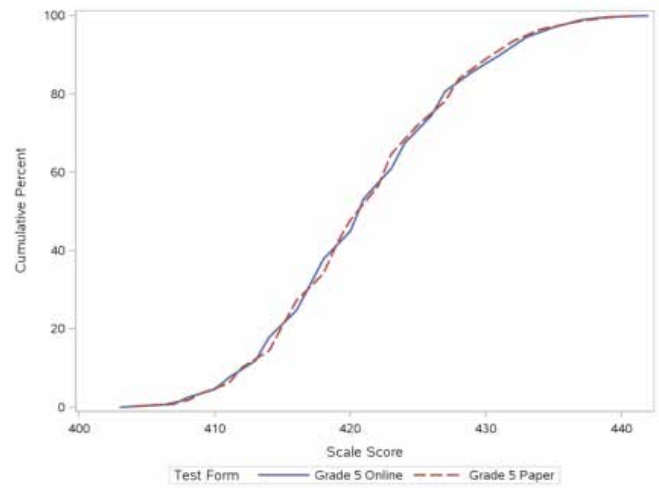
Grade 3



Grade 4



Grade 5



Grade 6

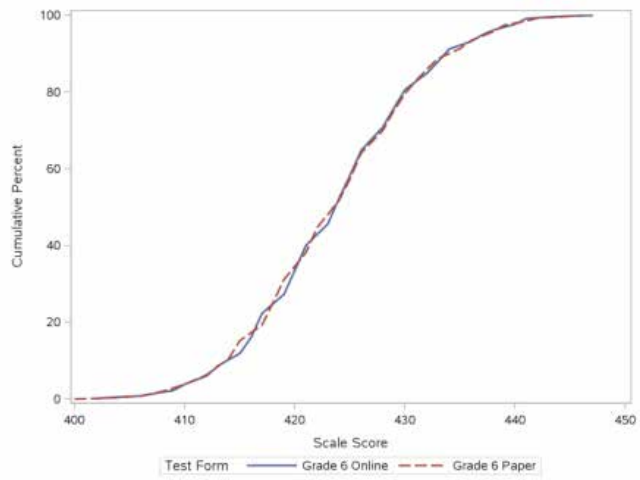
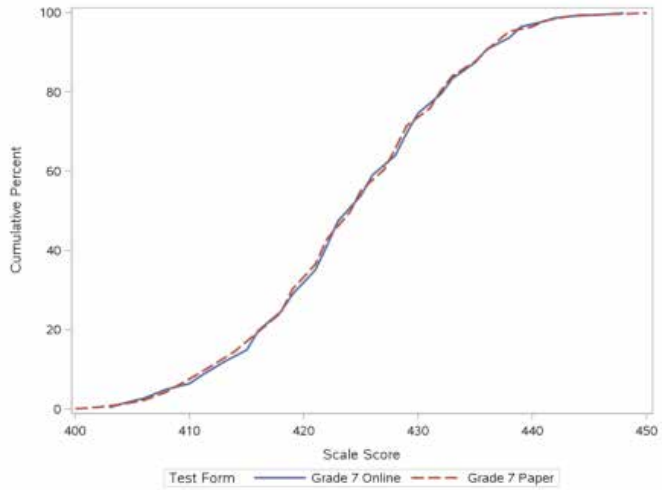
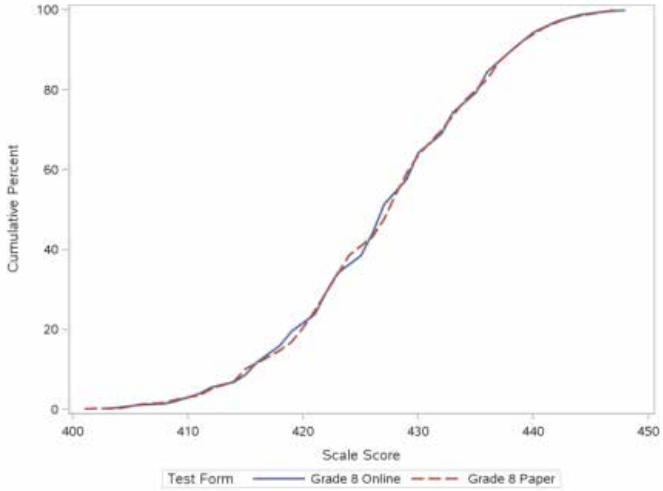


Figure 10.2. (continued)

Grade 7



Grade 8



Early High School

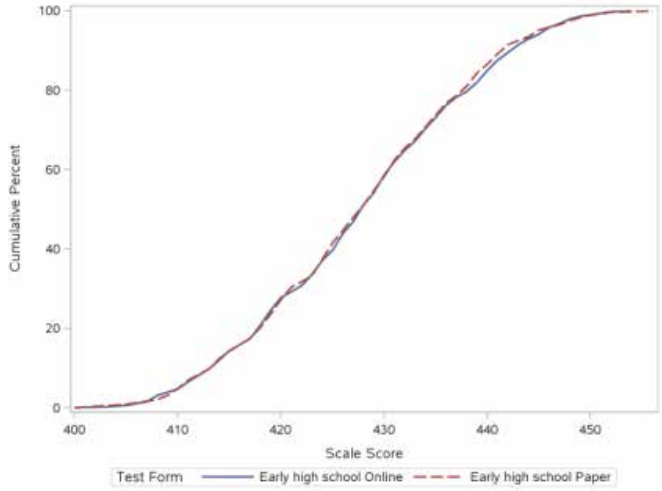
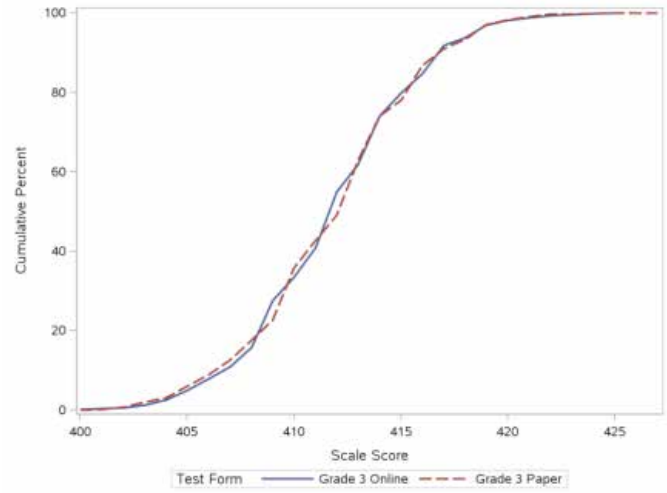


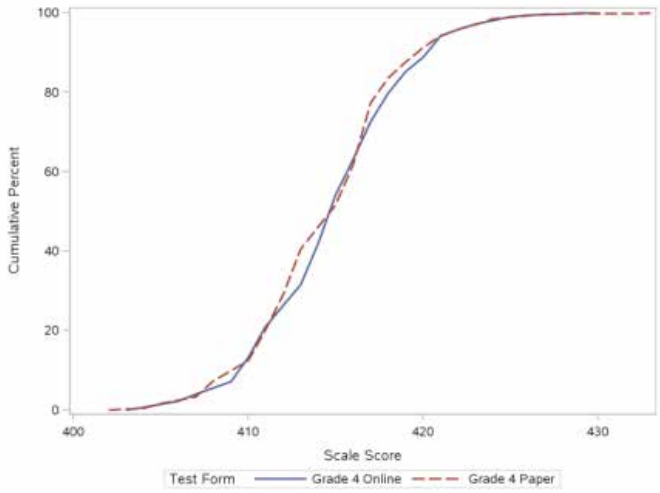
Figure 10.2. (continued)

Mathematics

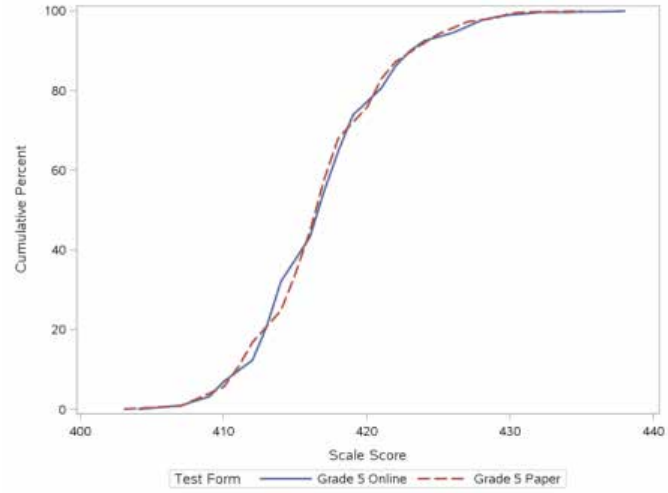
Grade 3



Grade 4



Grade 5



Grade 6

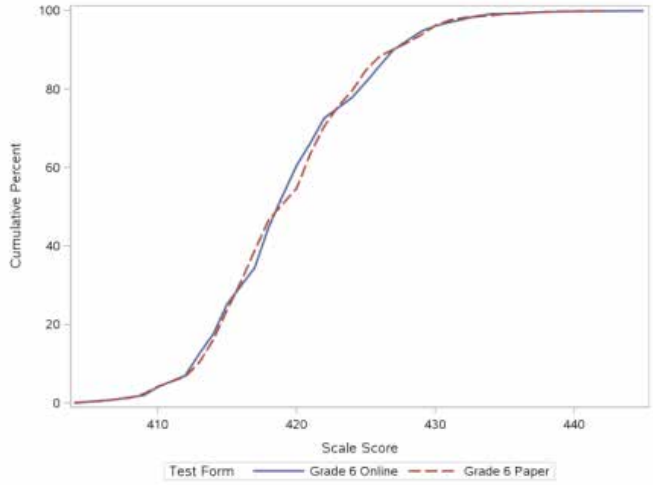
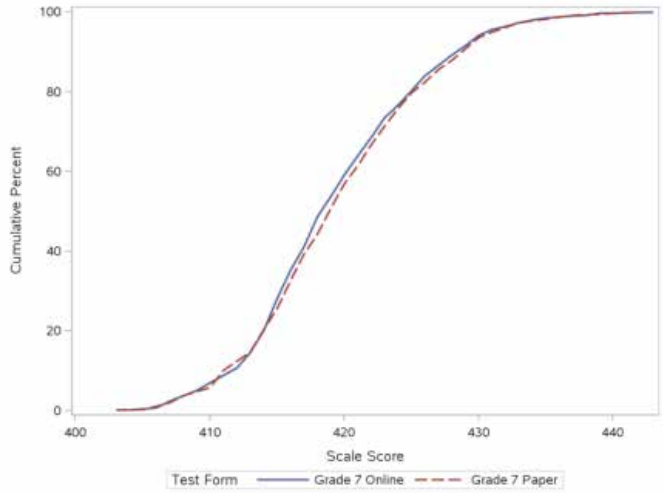
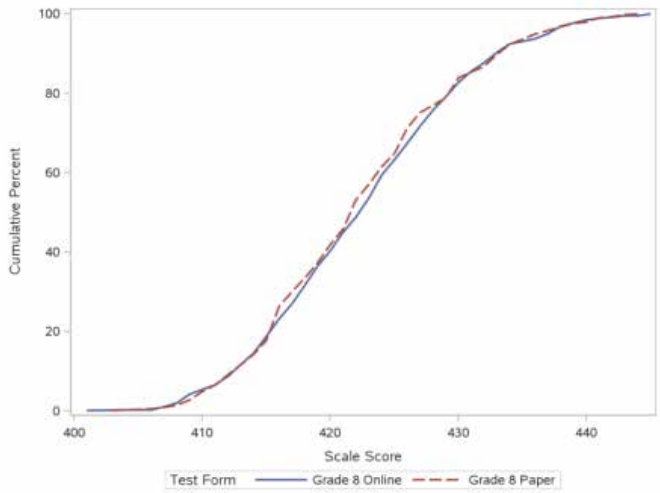


Figure 10.2. (continued)

Grade 7



Grade 8



Early High School

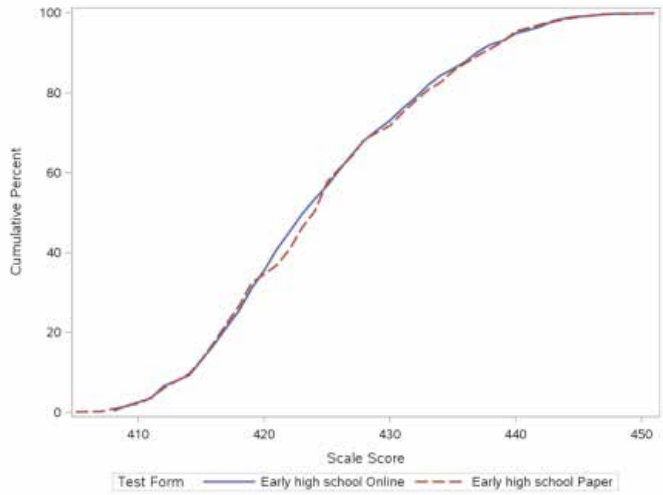
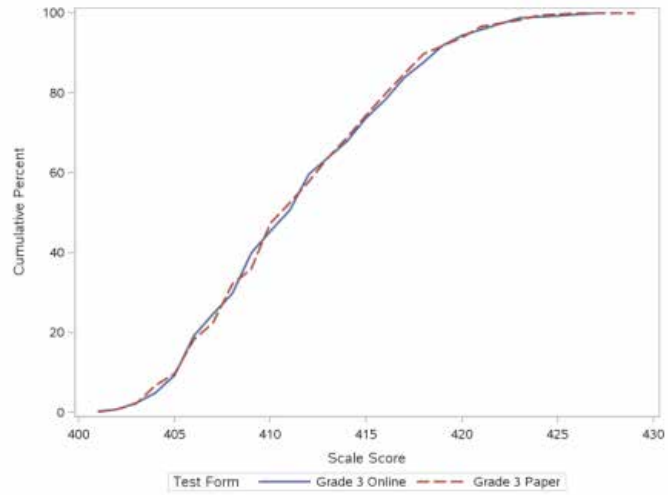


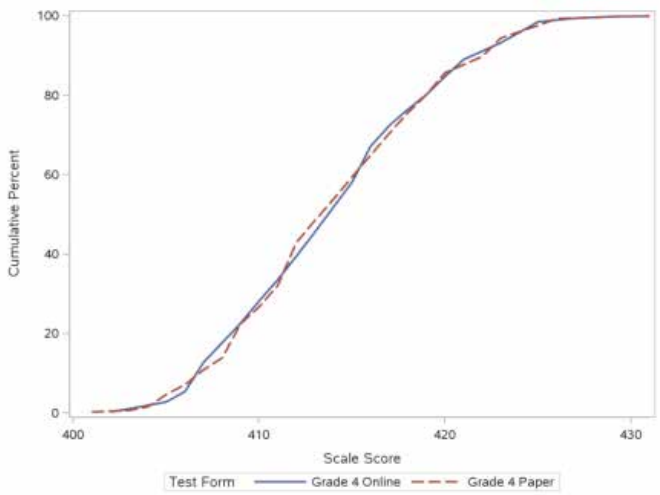
Figure 10.2. (continued)

Reading

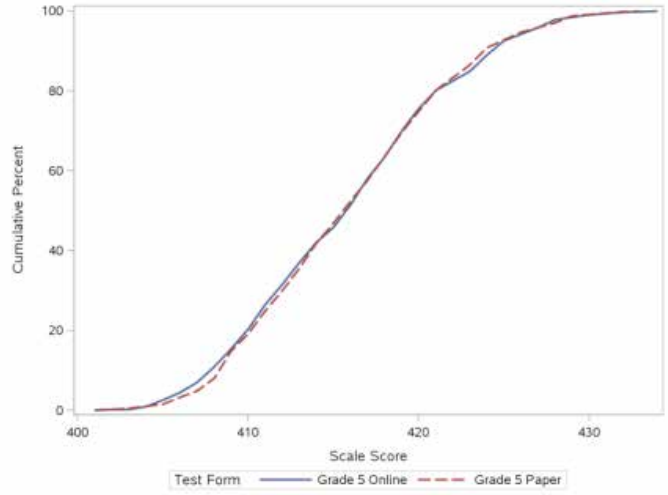
Grade 3



Grade 4



Grade 5



Grade 6

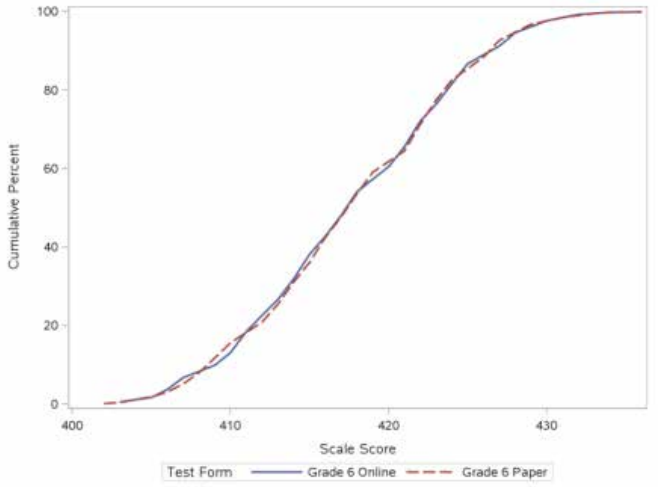
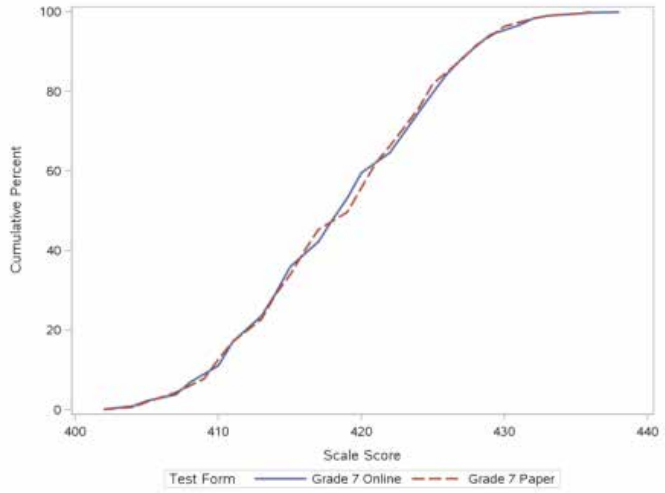
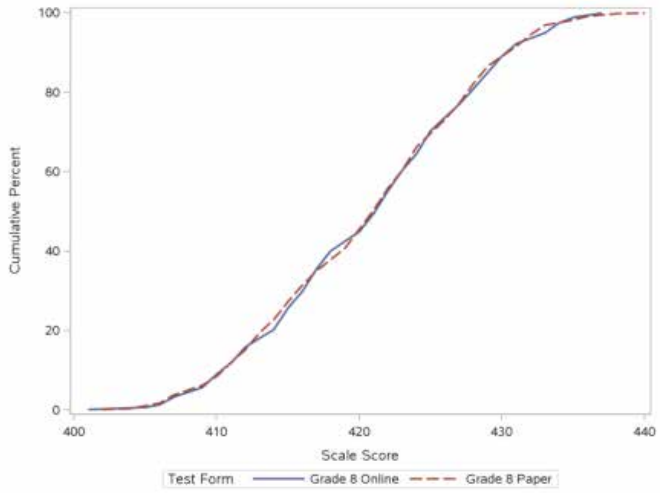


Figure 10.2. (continued)

Grade 7



Grade 8



Early High School

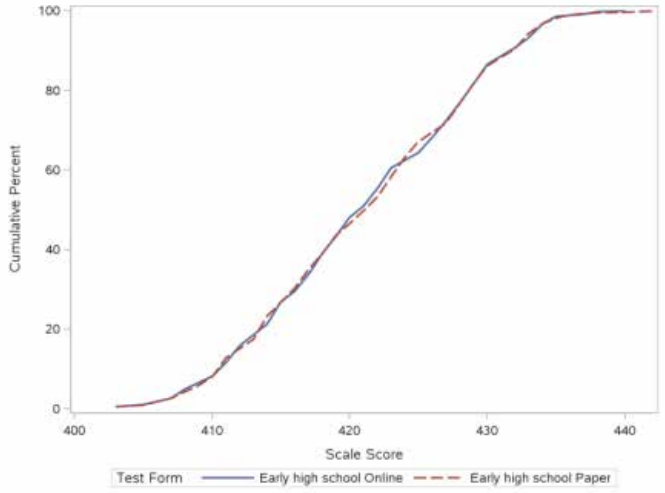
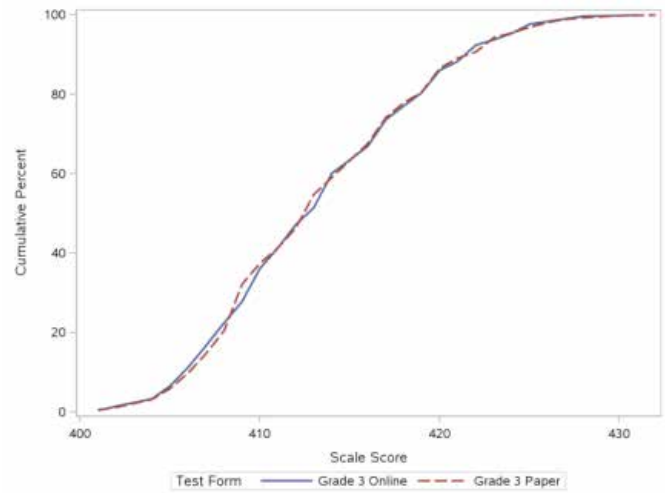


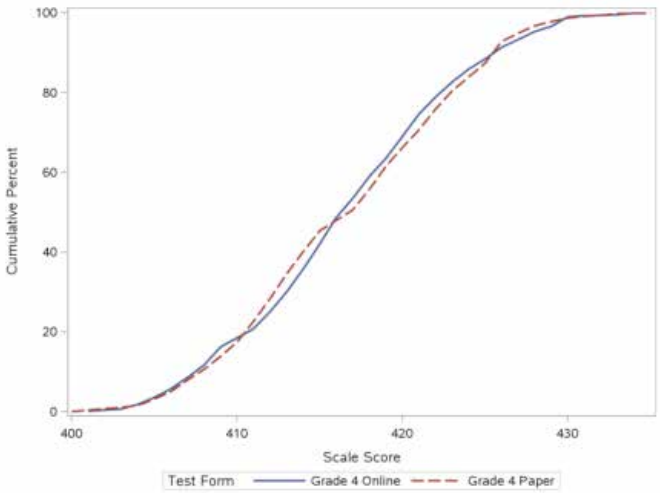
Figure 10.2. (continued)

Science

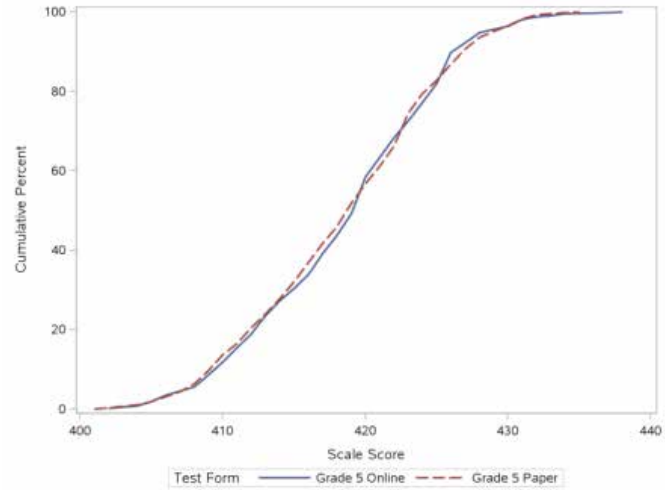
Grade 3



Grade 4



Grade 5



Grade 6

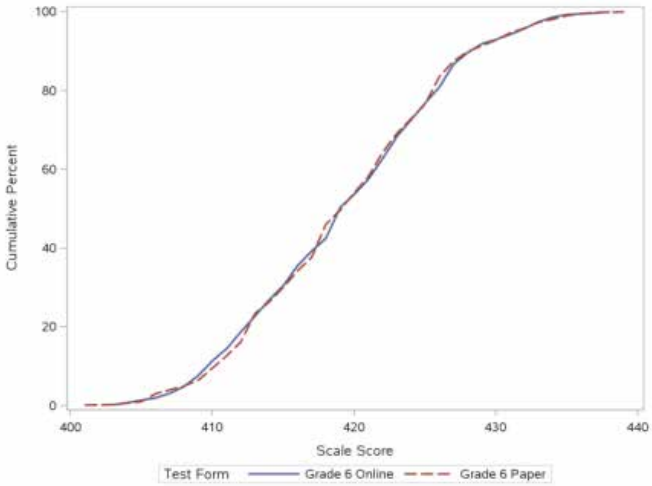
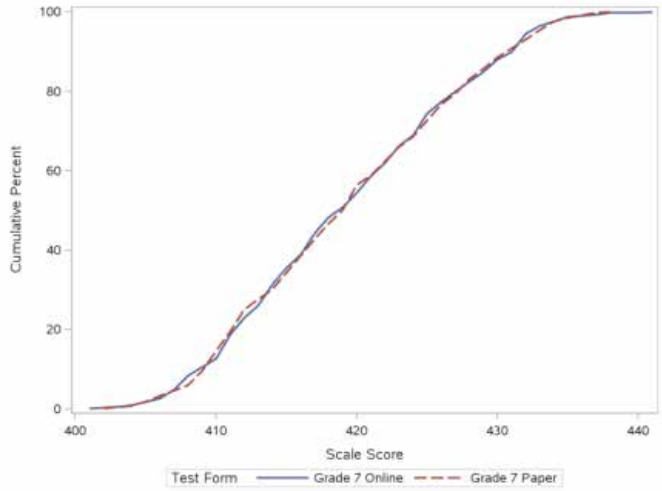
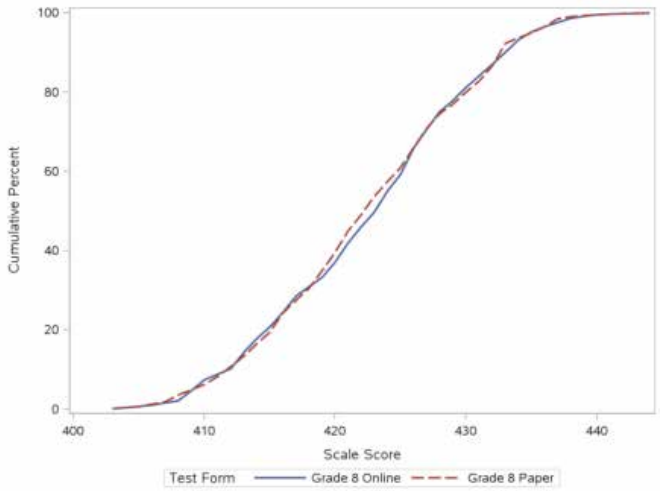


Figure 10.2. (continued)

Grade 7



Grade 8



Early High School

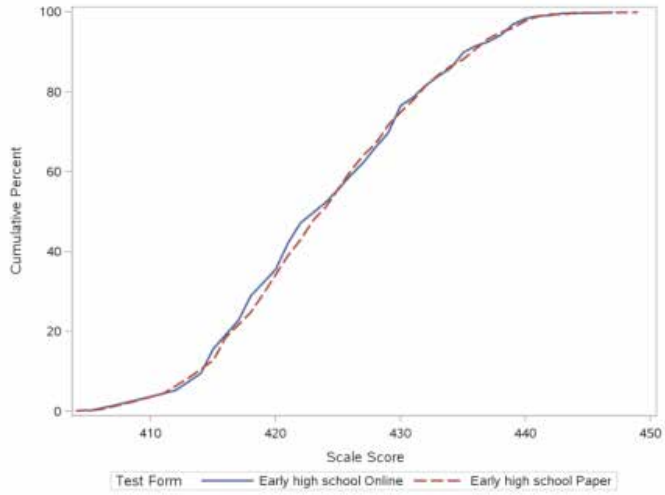
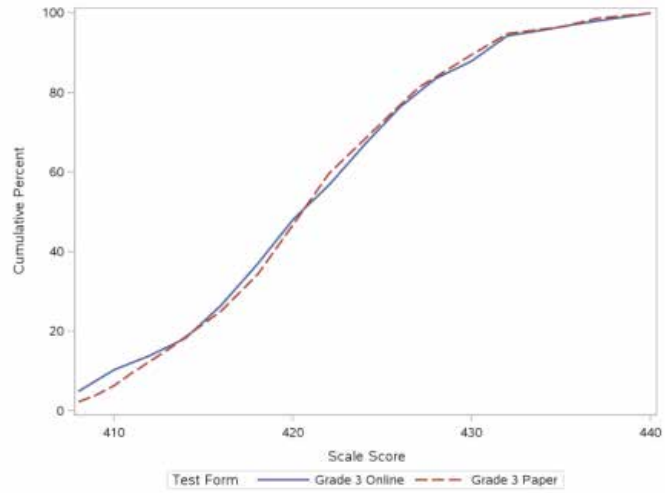


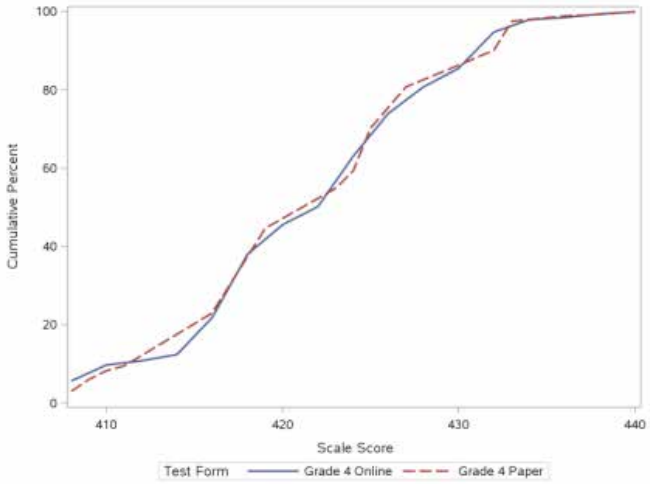
Figure 10.2. (continued)

Writing

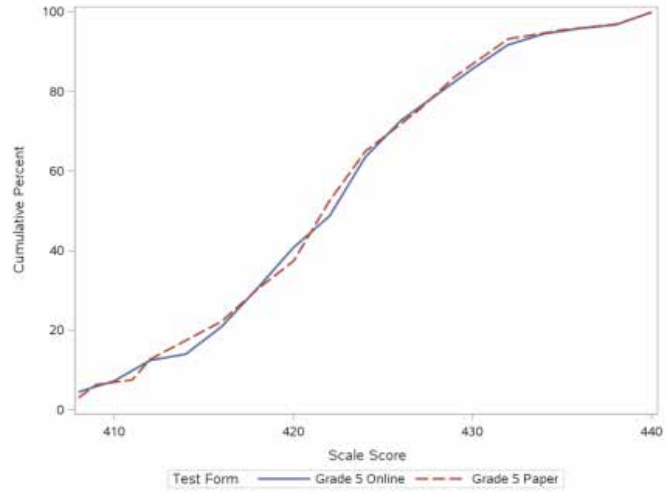
Grade 3



Grade 4



Grade 5



Grade 6

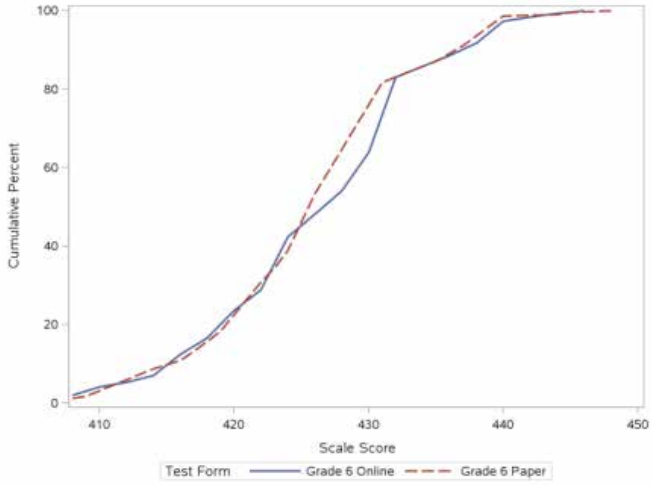
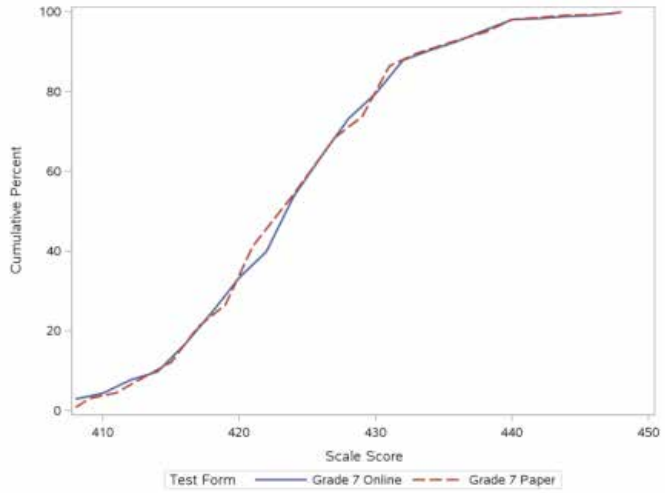
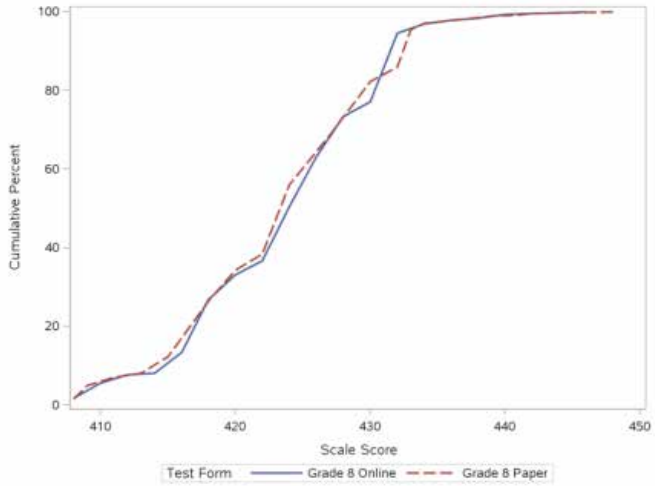


Figure 10.2. (continued)

Grade 7



Grade 8



Early High School

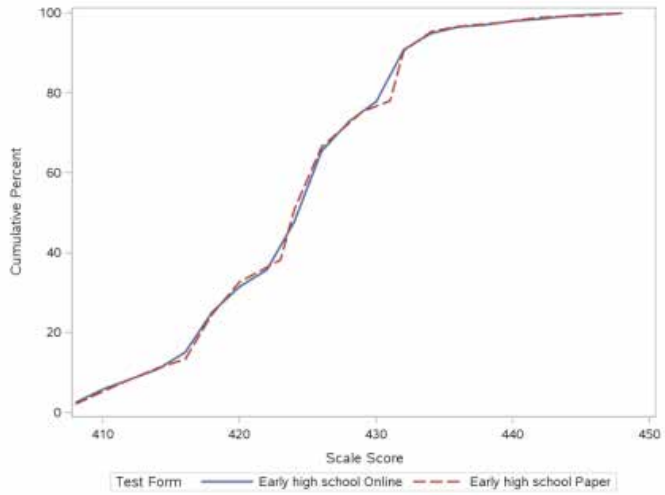
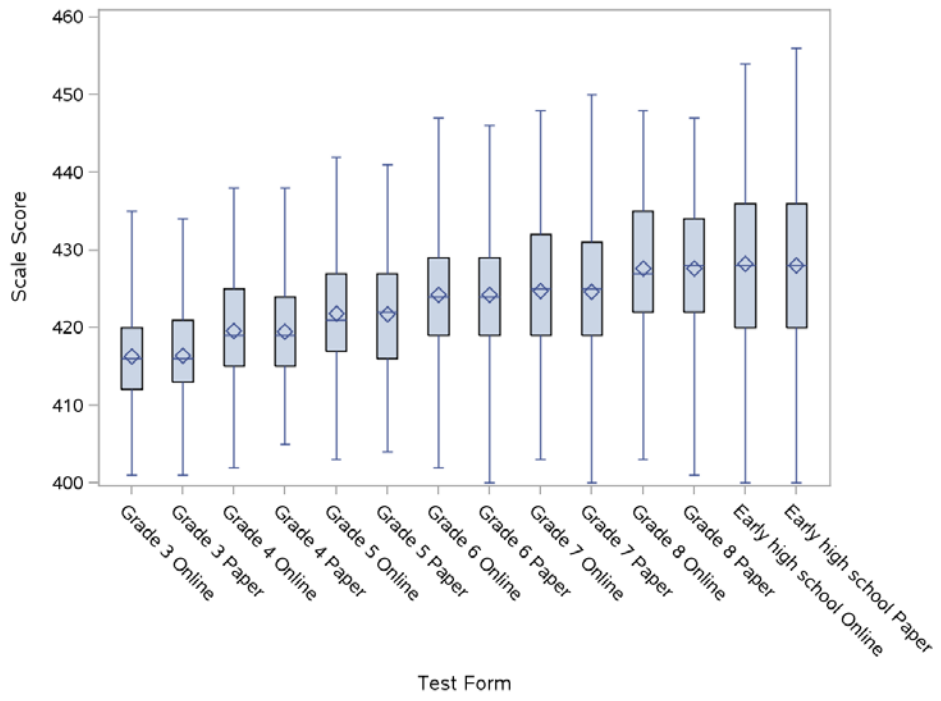


Figure 10.3. Box plots of scale scores by grade and mode for each subject area

English



Mathematics

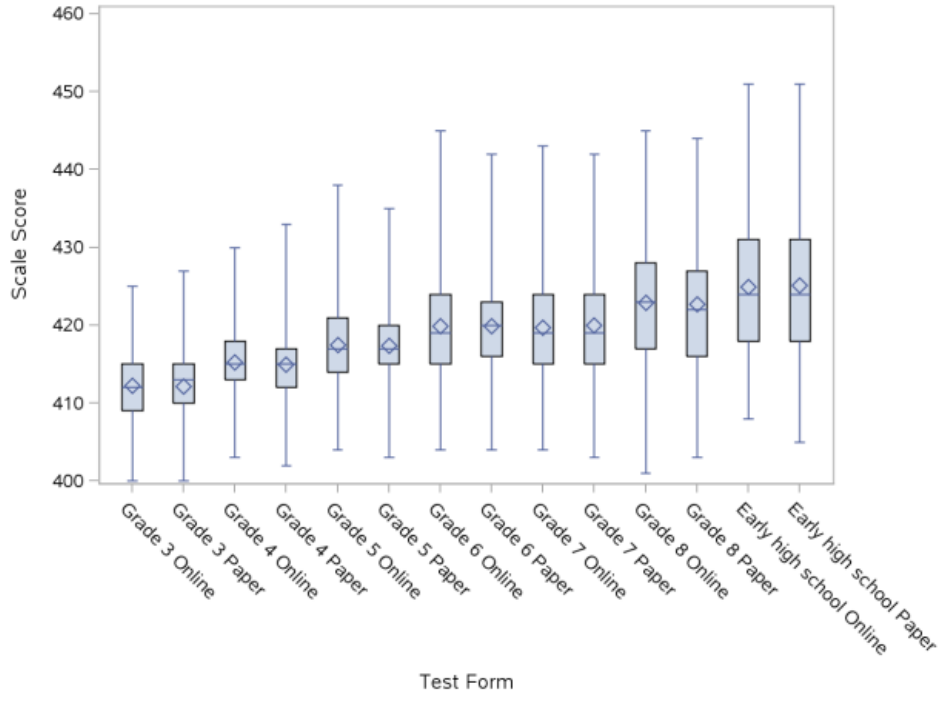
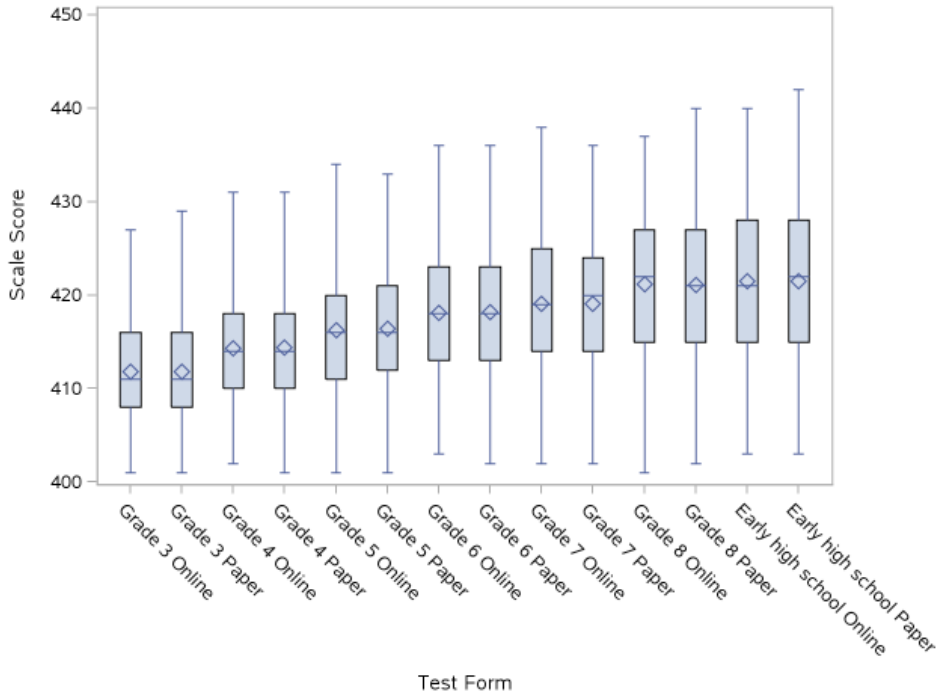


Figure 10.3. (continued)

Reading



Science

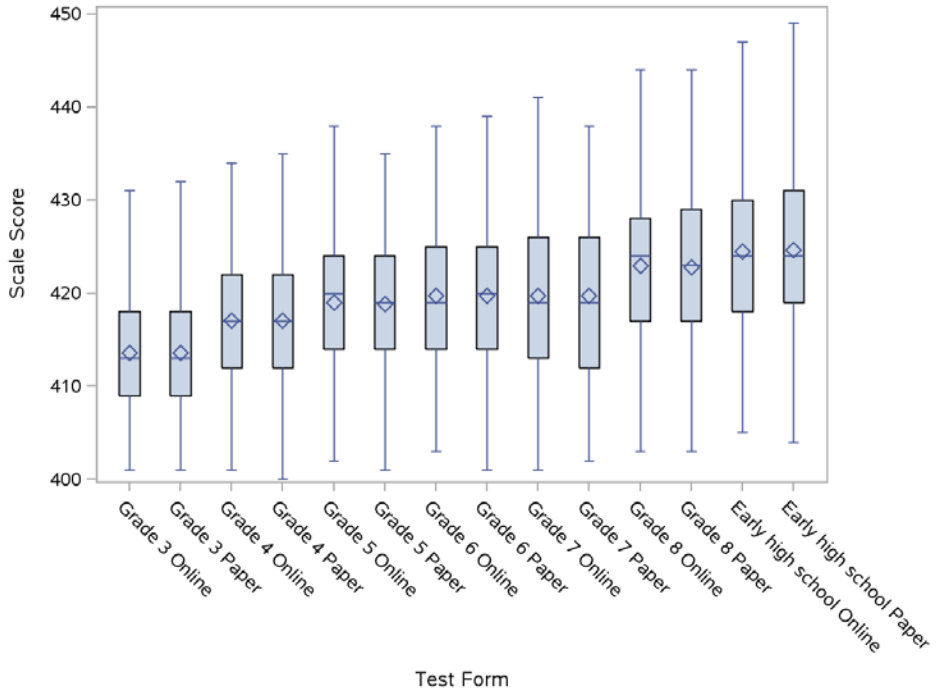
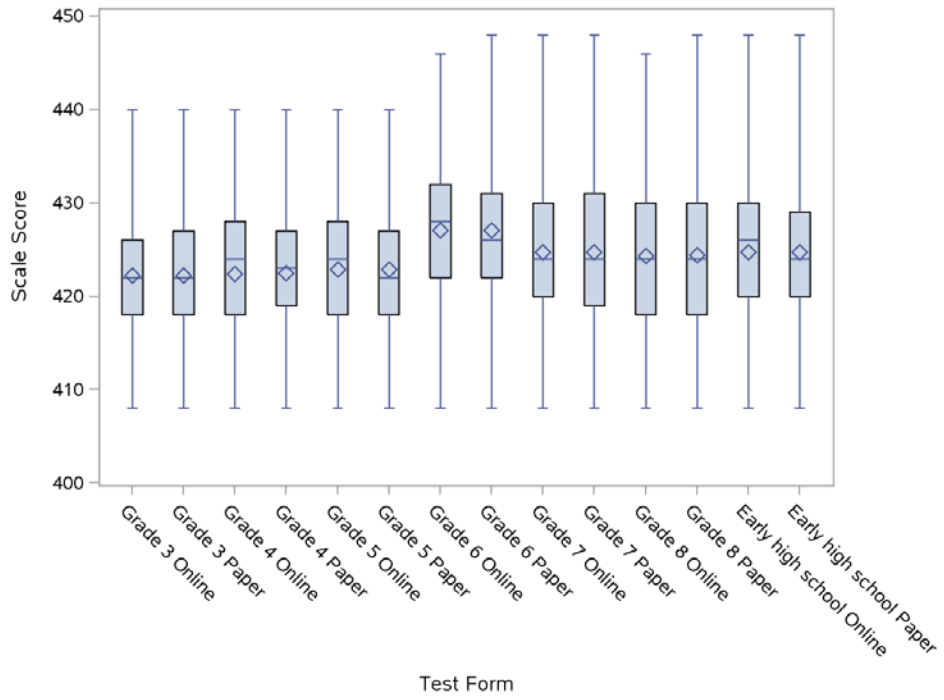


Figure 10.3. (continued)

Writing



Based in part on these results, we argue that the statistical linking was successful in assuring that scale scores across paper and online modes were comparable. In other words, mode effects across online and paper forms were effectively eliminated; scale scores appeared to function similarly across modes.

Summary

Based on the results comparing raw number-of-points-correct scores on identical items across modes, it is preferable to assume mode effects for ACT Aspire summative assessments. Not every grade and subject showed evidence of mode effects, but it occurred often enough that assuming no mode effects was not a reasonable assumption. When mode effects were observed, they generally appeared relatively small and in some cases (e.g., Writing) were inconsistent, if some degree of score differences due to mode were not a strong concern, it may be acceptable or even preferable to ignore mode effects on identical items across mode, but not under current interpretations of ACT Aspire scores.

As described above, some form of statistical linkage was required across modes due to item-type differences across modes. Statistical linkages appeared to successfully adjust for differences in items and mode across paper and online forms under the random groups design. Scale scores appeared comparable across mode.

CHAPTER 11

ACT Aspire Equating

Multiple ACT Aspire test forms are developed each year. Despite being constructed to follow the same content and statistical specifications, test forms may differ slightly in difficulty. Equating is used to control for these differences across forms so that scale scores reported to students have the same meaning regardless of the specific form administered. Equating is the process of making statistical adjustments to maintain score interchangeability across test forms (see Holland and Dorans 2006; Kolen and Brennan 2014).

ACT Aspire equating typically uses a random groups design, which involves spiraling the administration of test forms. In this case, test forms are interspersed within a classroom so that forms are distributed equally and randomly equivalent groups of students take each form. Under this design, if groups are indeed randomly equivalent, differences observed in performance across forms can be attributed to differences in form difficulty and equating methods applied to adjust for these differences.[20]

Each year, a carefully selected sample of students from an operational administration of ACT Aspire is used as an equating sample. Students in this sample are administered a spiraled set of ACT Aspire test forms that includes new test forms and an anchor test form that has already been equated to previous forms. Spiraling occurs separately for paper and online test forms, but a large sample of students takes each form.

[20] This methodology is also used to evaluate the comparability of forms when items are scrambled and when forms contain different pretest items.

Scores on alternate test forms are equated to the ACT Aspire score scale using equipercentile equating methodology (Kolen and Brennan 2014). In equipercentile equating, scores on different test forms are considered equivalent if they have the same percentile rank in a given group of students. Equipercentile equating is applied to the raw number-of-points scores for each subject test separately. The equipercentile equating results are subsequently smoothed using an analytic method described by Kolen (1984) to establish a smooth curve, and the equivalents are rounded to integers. The conversion tables that result from this process are used to transform raw scores on the new forms to scale scores.

In special cases where slight changes are made to the anchor form in the current year's equating study compared to its administration in the previous equating study, the revised anchor form is first equated to its original version using a common-item-nonequivalent group design, and then new forms are equated to the revised anchor using a random equivalent groups design.

Composite scores, including STEM and ELA scores, are not directly equated across forms but are calculated based on the separate subject test scale scores. Other scores, such as reporting category scores, are not equated across forms and are calculated based on raw number of points earned.

APPENDIX A

The ACT with Constructed-Response Items

While ACT Aspire covers grades 3 through 10, the scaling study also included grade 11. The primary reason for including grade 11 was so that the ACT Aspire vertical scale could be extended to a version of the ACT and provide scale continuity from grades 3 through 11. The implications of this are twofold. First, the scaling test (ST4) included ACT items so that the vertical scale could include grade 11 and be extended to a version of The ACT. Second, a grade 11 on-grade test was included in the scaling study.

The on-grade test covering grade 11 included in the scaling study had two parts: an operationally administered form of The ACT and a separately administered set of constructed-response-only items. These two components were combined and scaled to obtain scores on the ACT Aspire scale. This grade 11 test is referred to as *The ACT with constructed-response tests*. In some places, such as tables and figures with limited space, this test may be referred to as "ACT" or "The ACT."

The constructed-response component was included for Mathematics, Reading and Science Tests. Table A1 lists the number of constructed-response items included in each component. The ACT Writing Test was not included in the scaling study.

By including The ACT with constructed-response tests in the scaling study, the ACT Aspire vertical scale was extended to grade 11 and included The ACT combined with constructed-response items, consistent with tests for grades 3–10 that also included constructed-response items at grades 3–10.

Table A1. Number of Constructed-response Items and Number of Score Points Included in the On-Grade ACT with Constructed-Response Test Forms in the Scaling Study

Subject	Number of Constructed Response Items	Raw Number-of-Points Score Ranges
Mathematics	6	0–24
Reading	3	0–10
Science	11	0–30

APPENDIX B

EPAS to ACT Aspire Concordance

Table B1. EPAS to ACT Aspire Concordance

EPAS Scale Score	Concorded ACT Aspire Scale Score			
	English	Math	Reading	Science
1	400	400	400	400
2	401	402	401	401
3	402	403	402	402
4	403	404	403	403
5	404	406	404	403
6	405	407	405	404
7	406	408	406	405
8	408	409	407	406
9	410	409	408	407
10	414	411	410	408
11	417	412	411	409
12	419	414	413	410
13	422	415	415	412
14	424	417	418	414
15	426	420	420	416

Table B1. (continued)

EPAS Scale Score	Concorded ACT Aspire Scale Score			
	English	Math	Reading	Science
16	428	422	422	419
17	430	425	424	422
18	433	428	425	425
19	435	431	427	427
20	437	432	428	430
21	439	434	429	432
22	440	435	430	434
23	442	437	431	435
24	443	438	432	436
25	445	439	434	438
26	447	440	434	439
27	448	441	435	440
28	449	443	435	441
29	450	444	436	442
30	451	446	437	443
31	451	449	437	445
32	453	452	438	447

References

ACT. 2006. *Ready for College and Ready for Work: Same or Different*. Iowa City, IA: ACT. Retrieved from http://www.act.org/research/policymakers/pdf/ReadinessBrief.pdf.

———. 2007a. *National Career Readiness Certificate: WorkKeys Assessments Technical Bulletin.* Iowa City, IA: ACT.

———. 2007b. *The ACT Technical Manual.* Iowa City, IA: ACT.

———. 2013a. *The ACT Explore Technical Manual: 2013–2014.* Iowa City, IA: ACT.

———. 2013b. *The ACT Plan Technical Manual: 2013–2014.* Iowa City, IA: ACT.

———. 2014a. *ACT Aspire Exemplar Writing Questions.* Iowa City, IA: ACT.

———. 2014b. *ACT Aspire Summative Assessment Technical Bulletin #1.* Iowa City, IA: ACT.

Allen, Jeff. Forthcoming. *Development of Predicted Paths for ACT Aspire Score Reports.* Iowa City, IA: ACT.

Allen, Jeff, and Jim Sconing. 2005. *Using ACT Assessment Scores to Set Benchmarks for College Readiness.* ACT Research Report Series 2015-3. Iowa City, IA: ACT.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Baker, Frank B. and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques.* Second edition. New York: Dekker.

Betebenner, Damian W., Adam Van Iwaarden, Ben Domingue, and Yi Shang. 2014. *SGP: An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories.* R package version 1.2-0.0. http://cran.r-project.org/web/packages/SGP/index.html.

Campbell, Donald T. and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin 56 (2): 81–105.*

Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometricka* 16 (3): 297–334.

Cronbach, Lee J., Peter Schönemann, and Douglas McKie. 1965. "Alpha Coefficients for Stratified Parallel Tests." *Educational and Psychological Measurement* 25 (2): 291–312.

De Ayala, R. J. 2009. *The Theory and Practice of Item Response Theory. New York: Guilford.*

Geisinger, Kurt F. 1991. "Using Standard-Setting Data to Establish Cutoff Scores." *Educational Measurement: Issues and Practice* 10 (2): 17–22.

Gilmer, Jerry S. and Leonard S. Feldt. 1983. "Reliability Estimation for a Test with Parts of Unknown Length." *Psychometrika* 48 (1): 99–111.

Gulliksen, Harold. 1950. *Theory of mental tests. New York: Wiley.*

Hanson, B. A. 1994. *Extension of Lord-Wingersky Algorithm to Computing Test Score Distribution for Polytomous Items*. Unpublished Research Note.

Holland, Paul W. and Neil J. Dorans. 2006. "Linking and Equating." In *Educational Measurement*, edited by Robert L. Brennan. Fourth edition, 187–220. Westport, CT: American Council on Education and Praeger.

Kane, M. T. 2006. "Validation." In *Educational Measurement*, edited by Robert L. Brennan. Fourth edition, 17–64. Westport, CT: American Council on Education and Praeger.

Kingston, Neal M. 2009. "Comparability of Computer- and Paper-Administered Multiple-Choice Tests for K-12 Populations: A Synthesis." *Applied Measurement in Education* 22 (1): 22–37. doi: 10.1080/08957340802558326.

Koenker, Roger. 2005. *Quantile Regression.* New York, NY: Cambridge University Press.

Kolen, Michael J. 1984. "Effectiveness of Analytic Smoothing in Equipercentile Equating. *Journal of Educational Statistics* 9 (1): 25–44.

Kolen, Michael J. and Robert L. Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices*. Third edition. New York: Springer-Verlag.

Rogers, James, Kenneth I. Howard, and John T. Vessey. 1993. "Using Significance Tests to Evaluate Equivalence between Two Experimental Groups." *Psychological Bulletin* 113 (3): 553–565.

Serlin, Ronald C. and Daniel K. Lapsley. 1985. "Rationality in Psychological Research: The Good-Enough Principle." *American Psychologist* 40 (1): 73–83.

Stocking, Martha L. and Frederic M. Lord. 1983. "Developing a Common Metric in Item Response Theory." *Applied Psychological Measurement* 7 (2): 201–210.

Thurstone, Louis L. 1938. *Primary Mental Abilities*. Psychometric Monographs No. 1. Chicago: Univ. of Chicago Press.

Wang, Marilyn W. and Julian C Stanley. 1970. "Differential Weighting: A Review of Methods and Empirical Studies." *American Educational Research* 40 (5): 663-705.

Wang, Tianyou, Michael J. Kolen, and Deborah J. Harris. 2000. "Psychometric Properties of Scale Scores and Performance Levels for Performance Assessments Using Polytomous IRT." *Journal of Educational Measurement* 37 (2): 141–162

Wells, Craig S., Stephen G. Sireci, and Louise M. Bahry. 2014. *The Effect of Conditioning Years on the Reliability of SGPs.* Center for Educational Assessment Research Report Number 869. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst. http://www.umass.edu/remp/pdf/WSB_NCME2014.pdf

Williams, Valerie S. L., Mary Pommerich, and David Thissen. 1998. "A Comparison of Developmental Scales Based on Thurstone Methods and Item Response Theory." *Journal of Educational Measurement* 35 (2): 93–107.

Yen, W. M. and A. R. Fitzpatrick. 2006. "Item Response Theory." In *Educational Measurement*, edited by Robert L. Brennan. Fourth edition, 111–153. Westport, CT: American Council on Education and Praeger.