World-Class Instructional Design and Assessment

Annual Technical Report for ACCESS for ELLs® English Language Proficiency Test, Series 203, 2011-2012 Administration

Annual Technical Report No. 8
Volume 2 of 3: Analyses of Test Forms

Prepared by:

Mohammed Louguit, Ph.D.
Tiffany Yanosky, M.A.
Melissa Amos, M.S.
Shu Jing Yen, Ph.D.
Cathy Cameron, M.A.
David MacGregor, Ph.D.
Dorry M. Kenyon, Ph.D.

Center for Applied Linguistics

CONFIDENTIAL

June 3, 2013

Table of Contents

Volume 2

5. Analyses of Test Forms: Overview

This chapter contains two parts. The first part provides some background on the technical measurement and statistical tools used to analyze ACCESS for ELLs®. The second part explains the results that are presented for each test form in Chapter 6.

5.1 Background

5.1.1 Measurement Models Used

The measurement model that forms the basis of the analysis for the development of ACCESS for ELLs® is the Rasch measurement model (Wright and Stone, 1979). Additional information on its use in the development of the test is available in WIDA Technical Report 1, *Development and Field Test of ACCESS for ELLs®*. The test was developed using Rasch measurement principles, and in that sense the Rasch model guided all decisions throughout the development of the assessment and was not just a tool for the statistical analysis of the data. Thus, for example, data based on Rasch fit statistics guided the inclusion, revision, or deletion of items during the development and field testing of the test forms, and will continue to guide the refinement and further development of the test.

For Listening, Reading, and Speaking, the dichotomous Rasch model was used as the measurement model. Mathematically, the measurement model may be presented as

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) = B_n - D_i$$

where

P_{ni1} = probability of a correct response "1" by person "n" on item "i"

P_{ni0} = probability of an incorrect response "0" by person "n" on item "i"

B_n = ability of person "n"

D_i = difficulty of item "i"

When the probability of a person getting a correct answer equals the probability of a person getting an incorrect answer (i.e., 50% probability of getting it right and 50% probability of getting it wrong), P_{ni1}/P_{ni0} is equal to 1. The log of 1 is 0. This is the point at which a person's ability equals the difficulty of an item. For example, a person whose ability is 1.56 on the Rasch logit scale encountering an item whose difficulty is 1.56 on the Rasch logit scale would have a 50% probability of answering that question correctly.

For the Writing tasks, a Rasch Rating Scale model was used. Mathematically, this can be represented as

$$\log\left(\frac{P_{nik}}{P_{nik-1}}\right) = B_n - D_i - F_k$$

where

P_{nik} = probability of person "n" on task "i" receiving a rating at level "k" on the rating scale

P_{nik-1} = probability of person "n" on task "i" receiving a rating at level "k - 1" on the rating scale (i.e., the next lowest rating)

B_n = ability of person "n"

D_i = difficulty of task "i"

F_k = calibration of step "k" on the rating scale

All Rasch analyses were conducted using the Rasch measurement software program *Winsteps* (Linacre, 2006). Rasch statistics are presented in several of the tables that follow. When speaking of the measure of examinee ability, we use the term "ability measure" (rather than *theta* used commonly when discussing models based on Item Response Theory). When speaking of the measure of how hard an item was, we use the term "item difficulty measure" (rather than the *b parameter* used commonly when discussing models based on IRT). "Step measures" refer to the calibration of the steps in the Rasch Rating Scale model presented above. All three measures (ability, difficulty, and step) are expressed in terms of Rasch logits, which then are converted into scores on the ACCESS for ELLs® score scale for reporting purposes (see Technical Report 1 for more details).

Rasch model standard errors also appear in the tables. These are an indication of the precision with which the measures have been estimated. Unlike the Standard Error of Measurement (SEM) based on classical test theory, which posits the same SEM for all persons, regardless of where on the ability distribution they are, Rasch model standard errors are conditional on the individual's ability measure. All things being equal, if a person gets few items correct or few items incorrect, the standard error of that person's measure will be greater than if a person gets a moderate number of items correct. In addition, for ability measures, standard errors are a function of the number of items on a test form as well as the distribution and quality of the items (i.e., their fit to the Rasch model).

Also included in some of the tables are fit statistics for the Rasch model. These statistics are calculated by comparing the observed empirical data with the data that would be expected to be produced by the Rasch model. Of the several statistics available, the mean square fit statistics were used to flag items in the development of ACCESS for ELLs® that needed to be deleted or revised and are presented in the appropriate tables. Outfit mean square statistics are influenced by outliers. For example, a difficult item that for some reason some low - ability examinees get correct will have a high outfit mean square statistic that indicates that the item may not be measuring the same thing as other items on the test. Infit mean square statistics are influenced by more aberrant response patterns and generally indicate a more serious measurement problem. The expectation for both of these statistics is 1.00 and values near 1.00 are not of great concern. Values less than 1.00 indicate that the observations are too predictable and thus redundant, but are not of great concern. High values are more of a concern.

Linacre (2002, Autumn), the author of the Winsteps program, provides more guidance on how to interpret these statistics for test items. He writes:

- values greater than 2.0 “distort or degrade the measurement system”
- values between 1.5 and 2.0 are “unproductive for construction of measurement, but not degrading”
- values between 0.5 and 1.5 should be considered “productive for measurement”
- values below 0.5 Linacre calls “less productive for measurement, but not degrading”

Linacre also states in this guidance that infit problems are more serious to the construction of measurement than are outfit problems.

Because conservative guidelines were followed in the development of ACCESS for ELLs®, the vast majority of items and tasks on the test forms have mean square fit statistics in the range of 0.75 and 1.25, and fit the range that is “productive for measurement” according to the guidelines above.

5.1.2 Sampling

The results presented in most of the tables in Chapter 6 are based on the full data set of all students who were administered operational Series 203 of ACCESS for ELLs® in the academic year 2011-2012. Exceptions are Tables E, G, H, and I. The equating summary tables (Table E) use data from a sample of about 1,000 students rather than the entire population of students, because the equating was done in the midst of the operational scoring. The item or task analysis summary tables (Table G), the complete item analysis tables (Table H), and the raw score to scale score conversion tables (Table I) use item and task difficulties from this equating.

5.1.3 Equating and Scaling

Complete information on the horizontal and vertical scaling of ACCESS for ELLs® scores is provided in Technical Report 1, *Development and Field Test of ACCESS for ELLs®*. In brief, this scaling was accomplished during the field test based on an elaborate common item design, both across tiers and across grade-level clusters, which spanned two series of complete test forms. Concurrent calibration was used to determine item difficulty measures. These item difficulty measures were used to create the ACCESS for ELLs® scale scores used for reporting results on the test. Table D in Chapter 6 for each form provides the equation for converting Rasch ability measures in logits to ACCESS for ELLs® scale scores.

The operational test forms in Series 203 represent a partial refreshment of Series 202. That is, while many items were common on both forms, certain folders on Series 202 were replaced with new items. Thus, to place results on Series 203 onto the ACCESS for ELLs® score scale, items that were not revised or otherwise changed were anchored to the difficulty values from Series 202, which itself had been anchored to Series 201. Table E in Chapter 6 for each test form provides explicit information on the anchor items used for equating Series 203 results to those of Series 202.

5.1.4 DIF Analyses

Differential item analyses (DIF) attempt to investigate whether performances on items were influenced by factors extraneous to English language proficiency (i.e., the construct being measured on the test). In other words, it attempts to find items that may be functioning

differently for different groups based on criteria irrelevant to what is being tested. The performance of students on the ACCESS for ELLs® items was compared by dividing students into two different groupings: first, males versus females; second, students of Hispanic ethnic background versus students of all other ethnic backgrounds. (For both analyses, students for whom gender or ethnicity was missing were excluded.) Two commonly used procedures for detecting DIF were used: one for dichotomously scored items (Listening, Reading, and Speaking) and one for polytomously scored items (Writing).

5.1.4.1 Dichotomous Items

Following procedures originally proposed by the Educational Testing Service (ETS), the Mantel-Haenszel Chi-square statistic was used for dichotomous items. This procedure compares item-level performances of students in the two groups (e.g., males versus females) who are divided into subgroups based on their performance on the total test. It is assumed that, if there is no DIF, at any ability level (based on performance on the total test), a similar percentage of students in each group should get the item correct. The Mantel-Haenszel Chi-square statistic is used to check the probability that the two groups performed the same on each item across the ability groupings. The Mantel-Haenszel procedure is sensitive to detecting uniform DIF. The statistic is transformed into a scale called the "M-H delta" scale. This scale is symmetrical around zero, with a delta zero interpreted as indicating that neither group is favored. A positive result indicates that one group is favored; a negative result indicates that the other group is favored.

Because DIF is measured on a continuous scale, and because most items are likely to show some degree of DIF, it is useful to have guidelines to determine when the level of DIF is worrying. We follow the guidance provided by ETS to classify items into DIF levels as follows:

- A (no DIF), when the absolute value of delta was less than 1.0
- B (weak DIF), when the absolute value of delta was between 1.0 and 1.5
- C (strong DIF), when the absolute value of the delta was greater than 1.5

The software program *EZDIF* (Waller, n.d.) was used to run the DIF analyses for all forms containing dichotomous items. For each test form, the greatest number of ability level groupings is used; however, for many test forms, students scoring some of the lowest and highest raw scores need to be grouped together in order to have enough cases in each cell for the statistics to be appropriately calculated. (Note that this software program uses a two-step purification process; that is, items with C-level DIF in the first pass are removed from the matching variable in the second stage, and the DIF is then recalculated for the remaining items.)

(For information on procedures for dealing with items with C-level DIF, see Chapter 1.4.5.)

5.1.4.2 Polytomous Items

For polytomous items (i.e., the Writing tasks), a similar approach is used. It is based on the Mantel Chi-square statistic and the standardized mean difference following procedures again developed by ETS. As with dichotomous items, the underlying assumption is that students who performed similarly overall on the test should perform similarly on the individual tasks. To test this assumption, students are placed into six groups based on their total raw score on the Writing test. We determined these groups by calculating what the total raw score of a student scoring WIDA proficiency levels 1, 2, 3, 4, 5, or 6 in each groups would be. For example, a student

consistently scoring "1" would have a total score of "18" on a Tier B or Tier C form. A student consistently scoring "2" would score a "36."

To divide the students into performance groups in this way, cut points were made halfway between the above totals, such that students in Group 1 would have a total score of 0 to 27; Group 2 totaled 28 to 45; Group 3 totaled 46 to 63; Group 4 totaled 64 to 81; and Group 5 totaled 82 to 108. (Note that Group 5 contained students scoring in the 6 range. These two groups were combined because there are so few students in that category.)

For each Writing task, performance was similarly categorized according to the scoring rubric. Thus, raw scores of 0 to 4 were category 1 (i.e., up to a score totaling 4, such as 2-1-1, which is a high 1 but not yet a 2); the raw scores of 5 to 7 were category 2; the raw scores of 8 to 10 were category 3; the raw scores of 11 to 13 were category 4; the raw scores of 14 to 16 were category 5; and the raw scores of 17 to 18 were category 6. (The only exception to this was Kindergarten Writing tasks, where there was much smaller spread of scores on the Writing tasks. In such cases, total raw scores were used to determine categories.)

Following formulae provided by Zwick, Donoghue, & Grima (1993), an Excel spreadsheet was programmed to take cross-tabulated data output by SPSS, calculate the Mantel statistic, and determine its probability of significance. This Mantel statistic gives an indication of the probability that observed differences are the result of chance but does not indicate how significant that difference is. To indicate how significant the difference is, the standardized mean difference (SMD) between the performances of the two groups being compared is calculated. The standardized mean difference compares the means of the two groups, adjusting for differences in the distribution of the two groups being compared across the values of the matching variable. To standardize the outcome, this difference is divided by the standard deviation (SD) of the item for the total group. This calculation is also programmed into the Excel spreadsheet.

Following guidance proposed by ETS, polytomously scaled items are classified into DIF levels as follows:

- AA (no DIF), when the Mantel Chi-square statistic is not significant; or, when it is significant and the absolute value of (SMD/SD) is less than or equal to .17
- BB (weak DIF), when the Mantel Chi-square statistic is significant and the absolute value of (SMD/SD) is greater than .17 but less than or equal to .25
- CC (strong DIF), when the Mantel Chi-square statistic is significant and the absolute value of (SMD/SD) is greater than .25

5.2 Descriptions

The following paragraphs describe the tables that follow in Chapter 6 and are repeated for each test form in each domain.

5.2.1 Raw Score Information (Figure A and Table A)

Figure A and Table A relate to the *raw scores* on each test form. Listening, Reading, and Speaking were scored dichotomously (i.e., right or wrong). Thus, the highest possible score was the number of items on the test form. Each Writing task, however, could be awarded up to 18 points. Additionally, certain Writing tasks are weighted because of their potential to elicit higher levels of writing ability. For cluster 1-2, Tier A has a weight of 3 for the fourth task. For clusters 1-2, 3-5, 6-8, and 9-12, Tiers B and C have a weight of 2 for the second task and a weight of 3 for the third task. Thus, the maximum number of points on each Writing test form varies from 54 for the Tier A forms for clusters 3-5, 6-8, and 9-12 to 108 for the Tier B and C forms and cluster 1-2 Tier A.

For each test form, Figure A shows the distribution of the raw scores. The horizontal axis shows the raw scores. The vertical axis shows the number of students (count). Each bar shows how many students were awarded each raw score.

Table A shows, by each grade in the cluster and by total for the cluster:

- The number of students in the analyses (the number of students who were not absent, invalid, refused, exempt, or in the wrong cluster)
- The minimum observed raw score
- The maximum observed raw score
- The mean (average) raw score
- The standard deviation (std. dev.) of the raw scores

5.2.2 Scale Score Information (Figure B and Table B)

Figure B and Table B relate to the ACCESS for ELLs® *scale scores* on each test form. For each test form, raw scores were converted to vertically-equated scale scores. (The raw score to scale score conversion table for each test form is given as the last table—Table I—in each section.)

Thus, for each test form, Figure B shows the distribution of the scale scores. The horizontal axis shows the scale scores based on performances on the test form. To provide full perspective, it extends somewhat below and above the range of possible or observed scale scores. The vertical axis shows the number of students (count). Each bar shows how many students were awarded each scale score.

Table B shows, by each grade in the cluster and by total for the cluster:

- Number of students in the analyses (the number of students who were not absent, invalid, refused, exempt, or in the wrong cluster)
- The minimum observed scale score
- The maximum observed scale score

- The mean (average) scale score
- The standard deviation (std. dev.) of the scale scores

Note that scale scores for Tier A and Tier B in Listening and Reading are capped. Within each grade, the highest possible scale score for Tier A is the scale score corresponding to the cut score for proficiency level 4 (i.e., proficiency level score of 4.0). For Tier B, the highest possible scale score within each grade is the score corresponding to the cut score for proficiency level 5 (i.e., proficiency level score of 5.0). Because of the grade-level cut scores, the scale score associated with a given proficiency level score increases by grade level within a cluster, and so the cap also increases by grade level. For example, for 3-5A Listening, the scale score is capped at 325 for Grade 3, 338 for Grade 4, and 350 for Grade 5 (see Table 6.3.1.1B). Thus, a third grade student with a raw score of 20 (out of 20) on that test will have a scale score of 325, a fourth grader with the same raw score will have a scale score of 338, and a fifth grader with the same raw score will have a scale score of 350. However, all three students would have a proficiency level score of 4.0.

Also note that, because the scale is vertically equated, the range of scale scores moves up the scale from one cluster to the next. Thus, a second grade student with a raw score of 0 on the Tier A Listening test would have a scale score of 108, while a fifth grade student with a raw score of 0 on the Tier A Listening test would have a scale score of 120.

Similarly, scale scores at the lower end may be truncated so that the lowest achievable proficiency level score is 1.0. Again, this results in a lower minimum scale score for students in lower grade levels within a grade-level cluster.

The influence of these cuts will also be noticed in Figure B, as well as in many other tables throughout the report.

5.2.3 Proficiency Level Information (Figure C and Table C)

Figure C and Table C provide information on the proficiency level distribution of the students who took the test form. Thus, for each test form, Figure C shows the information graphically for the cluster as a whole. The horizontal axis shows the six WIDA proficiency levels. The vertical axis shows the percentage of students. Each bar shows the percent of students who were placed into each proficiency level in the domain.

Each row of Table C shows, by each grade in the cluster and by total for the cluster:

- The WIDA proficiency level designation (1 to 6)
- The number of students (count) whose performance on the test form placed them into that proficiency level in the domain being tested
- The percentage of students, out of the total number of students taking the form (by grade or by total for the cluster), who were placed into that proficiency level in the domain being tested

(Note that for some domains for Kindergarten and Tier A tests, it was not possible to place into all proficiency levels. Figure C and Table C also clearly show the effect of the scoring cap on Tiers A and B.)

For Kindergarten this information is provided for scores based on both the Accountability cut scores and the Instructional cut scores.

5.2.4 Scaling Equation Table (Table D)

For each test form, Table D provides the scaling equation for that domain. This equation is used to convert an examinee's ability measure into the scale score. Because ACCESS for ELLs® is vertically equated (see Chapter 5.1.3 above), though each domain has its own equation, the same equation is used across all tiers and grade-level clusters within each domain.

5.2.5 Equating Summary (Table E)

Each year a certain percentage of items on each ACCESS for ELLs® test form are refreshed. A post-equating procedure known as common item equating is used to equate results on new forms to the older forms. This means that the difficulty measure of items appearing on the new form that are the same as those on the older form are kept constant across both forms. Thus, performances on the newer form may be interpreted in the same frame of reference.

Many items appearing on ACCESS for ELLs® Series 203 also appeared on Series 202. All items common to both forms were anchored in the first equating run. After the first equating run, some items that were originally anchored proved to have changed in their difficulty measure. This change is measured by the "Displacement" statistic. This statistic shows the difference between the difficulty value of the anchored item and what its difficulty value would have been had it not been anchored. For Listening and Reading items and for Writing and Speaking tasks, if this value was large (i.e., usually above .30 or below -.30), that item was unanchored in the final equating run (i.e., it was treated as if it were a new item).

Table E presents a summary of the common item equating procedures. The first section of the table compares the current test (i.e., the Series 203 version of that test form) to the previous year's test (i.e., the Series 202 version of that test form). The number of items, the average item difficulty, the standard deviation of the item difficulty values, as well as the difficulty value of the easiest and hardest item on each test form is presented. These values are in terms of logits used in the Rasch measurement model.

The second section of the table presents information on the anchoring items. The total number of possible anchors (i.e., all common items) is shown, as well as the average difficulty and standard deviation of those items. Next, the number of items that were actually anchored (i.e., in general, those items whose displacement values were below .30 or above -.30) in the final equating run is shown, again with the average item difficulty and standard deviation. Finally, the percentage of items that served as anchors and the average displacement value is given. Generally speaking, the greater the number of tasks anchored and the closer the average displacement is to 0.00, the more trustworthy the equating results will be.

The third section of Table E shows the location of the anchor items or tasks, both by order on the test form and by order of difficulty. It is desirable that the anchored items appear throughout the test form in order to ensure that no systematic bias affects performance on them (e.g., if they all appear at the end of a test form, there may be a fatigue effect). It is also desirable that the anchor items represent a wide range of difficulties across the entire spectrum of the item difficulty values on a test form. The greater the representation across the difficulty range, the more trustworthy the equating results will be. This section also provides information on displacement; that is, the difference between the difficulty value of the anchored item and what that difficulty

value would have been had the item not been anchored. Smaller displacement statistics indicate more consistency between the item's difficulty value on the Series 203 test form and on the Series 202 test form. Typically, random displacements of less than 0.5 logits are unlikely to have much impact on measurement in a test instrument (Displacement measures, 2006, January 29).

Note that for the Writing tasks, this table also provides the anchored step measures for the total score on each task. For the ACCESS Writing tasks, a rating scale model is used (see Chapter 5.1.1 above). Because a single generic rubric based on the generic WIDA performance level definitions is used to score all of the Writing tasks across all of the grade-level clusters, we modeled a rating scale that has the same step difficulty values across all Writing tasks across all grade-level clusters. Thus, these values are the same for every Writing task on ACCESS. These constant step difficulty values help to provide anchors in the calibration of new Writing tasks onto the common WIDA score scale each year.

The step measure, or step difficulty, is the calibrated measure of the transition from the category below to the current category. It indicates how difficult it is to observe a category, not how difficult it is to perform it (Linacre, 1999). In November 2011, a study was conducted for the WIDA ACCESS TAC Meeting. This study examined whether disordered step measures were present in the field test data as well as in the Series 202 operational test data. For the field test, after the vertical reading scale was established, the Reading and Writing items were calibrated together across grade-level clusters, anchoring on the vertically-scaled Reading item parameters. Then, Writing task difficulties and scale step measures were anchored, and examinee abilities within each grade-level cluster on the vertical scale were determined. For the operational test, the task difficulty parameters and step measures were estimated by anchoring on students' ability measures. Analysis of the field test data and the operational data produced similar findings. For both of the sets of data, the main, or most commonly observed, raw scores on ACCESS (i.e., 0, 3, 6, 9, 12, 15, and 18) had scale step measures that increased. With one exception in the operational data, no disordering in the step measures of the main scores was found. In addition, for score points that were less commonly observed, disordered step measures were found, but were attributed to the nature of the scoring rubric. On the rubric, for example, it is much more common to attain score point 3 (1-1-1) than score point 4 (e.g., 2-1-1, 1-2-1, etc.). Thus, the overall conclusion of this study was that the disordering step measures did not indicate substantive problems with the rating scale definition.

Note that because the Kindergarten test form was newly created for Series 200, it was not equated to the Series 103 test. Therefore, Table E is not included for Kindergarten. For technical details on the Kindergarten test, see MacGregor, Kenyon, Gibson, and Evans (2009). In addition, in the other grade-level clusters, scores for the Speaking test are based on a content analysis rather than on equating to previous forms; therefore, Table E is included only to verify that the raw score to scale score conversion remains within reasonable parameters.

5.2.6 Test Characteristic Curve (Figure D)

For each test form, Figure D graphically shows the relationship between the ability measure (in logits) on the horizontal axis and the expected raw score on the vertical axis. Five vertical lines indicate the five cut scores for the highest grade in the cluster for the test form, dividing the figure into six sections for each of the WIDA proficiency levels (1–6) for the domain being tested. (Note that for some domains for Kindergarten and Tier A tests, it was not possible to

place into all six language proficiency levels.) As would be expected, higher raw scores are required to be placed into higher language proficiency levels. The relative width of each section between the cut score lines, however, gives an indication of how many items on that form must be answered correctly (or points on the Writing section must be earned) to be placed into a WIDA language proficiency level.

5.2.7 Test Information Function (Figure E)

With the Rasch measurement model, as with any measurement model following Item Response Theory (IRT), the relationship between the ability measure (in logits) and the accuracy of test scores can be modeled. It is recognized that tests measure most accurately when the abilities of the examinees and the difficulty of the items are most appropriate for each other. If a test is too difficult for an examinee (i.e., the examinee scores close to zero), or if the test is too easy for an examinee (i.e., the examinee "tops out"), accurate measurement of the examinee's ability cannot be made. The test information function shows graphically how well the test is measuring across the ability measure spectrum. High values indicate more accuracy in measurement. Thus, for each test form, Figure E shows the relationship between the ability measure (in logits) on the horizontal axis and measurement accuracy, represented as the Fisher information value (which is the inverse squared of the standard error), on the vertical axis. The test information function, then, reflects the conditional standard error of measurement.

Again, as in Figures D, five vertical lines in Figure E indicate the five cut scores for the highest grade in the cluster for the test form, dividing the figure into six sections for each of the WIDA language proficiency levels (1–6) for the domain being tested. (Note that for some domains for Kindergarten and Tier A tests, it was not possible to place into all six language proficiency levels. Note also that, although Listening and Reading scores on Tiers A and B were capped, all 5 horizontal lines indicating the cut points remain in this figure.) It is important that each test form measure most accurately in the areas for which it is primarily used to make classification decisions. In other words, optimally the test information function should be high for the cuts between 1/2 and 2/3 for Tier A test forms; between 2/3, 3/4, and 4/5 for Tier B test forms; and between 3/4, 4/5, and 5/6 for Tier C test forms.

5.2.8 Reliability (Table F)

In contrast to Figure E, which is based on the Rasch measurement model, Table F presents reliability and accuracy information based on Classical Test Theory. It shows:

- The number of students
- The number of items
- Cronbach's coefficient alpha (as a measure of internal consistency)
- The classical standard error of measurement (SEM) in terms of ***raw scores***

Cronbach's coefficient alpha is widely used as an estimate of reliability, particularly of the internal consistency of test items. It expresses how well the items on a test appear to measure the same construct. Conceptually, it may be thought of as the correlation obtained between performances on two halves of the test, if every possibility of dividing the test items in two were attempted. Thus, Cronbach's alpha may be low if some items are measuring something other

than what the majority of the items are measuring. As with any reliability index, it is affected by the number of test items (or test score points that may be awarded). That is, all things being equal, the greater the number of items, the higher the reliability.

Cronbach's alpha is also affected by the distribution of ability within the group of students tested. All things being equal, the greater the heterogeneity of abilities within the group of students tested (i.e., the more widely the scores are distributed), the higher the reliability. In this sense, Cronbach's alpha is *sample dependent.* It is widely recognized that reliability can be as much a function of the test as of the sample of students tested. That is, the exact same test can produce widely disparate reliability indices based on ability distribution of the group of students tested. Because ACCESS for ELLs® is a tiered test (that is, because each form in Tier A, B, or C targets only a certain range of the entire ability distribution), results for reliability on any one form may at times be lower than typically expected.

The formula for Cronbach's alpha is

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_t^2}\right]$$

where

n = number of items i

σ_i^2 = variance of score on item i

σ_t^2 = variance of total score

Table F also presents the *standard error of measurement* (SEM) based on classical test theory. Unlike IRT, in this approach, SEM is seen as a constant across the spread of test scores (ability continuum). Thus, it is ***not*** conditional on ability being measured. It is, however, a function of two statistics: the reliability of the test and the (observed) standard deviation of the test scores. It is calculated as

$$\text{SEM} = SD\sqrt{1 - reliability}$$

Traditionally, SEM has been used to create a band around an examinee's observed score, with the assertion in the view of classical test theory, that the examinee's true score (i.e., what the examinee's score would be if it could be measured without error) would lie with a certain degree of probability within this band. Statistically speaking, then, there is an expectation that an examinee's true score has a 68% probability of lying within the band extending from the observed score minus 1 SEM to the observed score plus 1 SEM.

For the Writing tests (except Kindergarten, which is scored by the test administrator), information on interrater reliability is also provided in Table F. This portion of the table shows, for each of the three or four Writing tasks, the percentage of agreement between two raters in terms of the three features being rated: Linguistic Complexity (LX), Vocabulary Usage (VU), and Language Control (LC). In this part of the table, the first column shows the Writing task (i.e.,

the first, second, third, or fourth, if applicable). The second column shows the number of Writing papers that were double scored. This number is generally 25% of all papers scored, chosen at random during the operational scoring process. The next column shows the feature, while the following columns show the rates of agreement: Exact, ADJ (Adjacent), and total Exact and Adjacent. When the two raters agreed on the score, an exact agreement was counted. If the two raters were different in that feature by one point, an adjacent agreement was counted.

All operational Speaking tests are scored by the test administrator. In this report, information on inter-rater reliability for Speaking provided in Table F (except for Kindergarten) is based on data from the field test of the Speaking test, reported on fully in ACCESS for ELLs® Technical Report 1, *Development and Field Test of ACCESS for ELLs®*. This portion of the table shows, for each of the 13 Speaking tasks, the number of individuals in the sample responding to the task, the number of agreements between two raters as to the rating of the task, and the percent agreement of the rating.

5.2.9 Item/Task Analysis Summary (Table G)

Table G provides a summary of the analyses of the items (for Listening and Reading) or the tasks (for Writing and Speaking). The top part of the table gives an item or task summary. The first column in this part states the type of item (MC for multiple choice or ECR for extended constructed response). The next column shows the number of items or tasks on the test form. The next column gives the average item or task difficulty value in logits. For the multiple-choice items, the next column shows the average p-value. This is the average percentage of correct items. The last two columns give information on the Rasch model fit statistics (see Chapter 5.1.1). The first is the average infit mean square statistic; and the second is the average outfit mean square statistic. Optimally, these values should be close to 1.00.

The next section of Table G provides a summary of the findings of the DIF analyses (see Chapter 5.1.4). The first column gives the DIF level: A, B, or C for dichotomous items or AA, BB, or CC for polytomous tasks (i.e., Writing tasks). The next major columns show the contrasting groups in the DIF analyses: either male versus female (M/F) or Hispanic versus other ethnicities (H/O). Even though DIF may be negligible (category A or AA), this table shows the number of items that were favoring one group or the other at all levels of DIF. Optimally, even when items are all in category A or AA, there should be roughly an even number of items favoring each of the two groups to ensure that there is no systematic biasing test effect across items.

For the Writing tasks, the last part of this table shows the distribution of the raw scores on each task by total score category. (Recall that the total score for a task equals the sum of three feature scores, which are scored from 1 to 6, for a maximum total of 18; however, papers that are written in languages other than English or are totally incomprehensible may receive a score of 0, while papers that demonstrate the ability to copy or write a few words in English may be awarded a score of 1. The total score of 2 is impossible to achieve.)

5.2.10 Complete Item Analysis Table (Table H)

Table H presents results of the analyses of all of the items or tasks on the test form. The first column provides a descriptive name of the item or task. The item or task names vary slightly across domains and grade-level clusters, but they usually consist of characters that represent the domain (e.g., "R" for Reading), the grade-level cluster (e.g., "g91" for grades 9–12), the tier (e.g., C, if applicable), the unique number in the item database (e.g., 3820), the WIDA Standard (e.g.,

"MA" for the Language of Mathematics), the language proficiency level targeted (e.g., "p3"), the thematic folder name (e.g., "Cafeteria"), and the test series (e.g., 203). Note that for Writing, "IT" stands for the "integrated" task, which requires more extensive writing and that integrates model performance indicators for SI, LA, and SS. Also, note that for some Kindergarten tasks, the naming system is a bit simpler, e.g., "1.S_A1_K_203", which contains the item order, domain, the folder, the proficiency level, the grade-level cluster, and the test series.

The second column in Table H presents the item difficulty in logits, while the third column indicates whether that item served as a common item, anchoring the measurement scale to the results of the field test. For dichotomously scored items (Listening, Reading, and Speaking), the fourth column shows the p-value (percent of correct answers on that item or, in the case of Speaking, percent of students meeting the expectations of that task). The next two columns show the Rasch fit statistics for the item or task, while the following columns show the results of the two DIF analyses for that item or task. These last columns are interpreted just as in Table G.

The speaking test consists of three thematic folders (A, B, and C). Each folder consists of three to five tasks written around a common theme. Generally speaking, tasks within folders were designed to illicit speaking proficiency levels from low to high with increasing demand or difficulty level as the folder progresses from the first to the last task. In addition, folders were designed to be more challenging from Folder A to C. Due to the semi adaptive nature of the test administration, when a task is scored "Approaches" or "No Response," the student moves to the next folder rather than continue with tasks in the current folder. The remaining tasks in that folder are marked as 'Not Administered,' which is converted to score of '0.' The Test Administration Manual (WIDA], 2011) provides more detailed information on the Speaking Test administration. The assumption is that the student would not have been able to meet the expectation of the remaining (more difficult) tasks in the folder.

In November 2011, a study was conducted for the WIDA ACCESS TAC Meeting. This study revealed that some large outfit statistics were caused either by non-conforming response patterns within a folder (i.e., a score of 0 for one task in a folder, and 1 for a later, more challenging task in that same folder) or by non-conforming response patterns across folders (i.e., the examinee does not meet expectations of any or some tasks in a less challenging folder but meets expectations of all or most of the tasks in another more challenging folder).

The "within folder non-conforming" responses appear to be to the result of either administration or recording errors because the response pattern is so unexpected. For "across folder non-conforming" responses, such as '000,1111' in Folders A and B, however, we do not have information to confirm or disconfirm that these unlikely response patterns are due to administrator or recording error.

Since the Series 201 Speaking test, CAL has removed the "within folder non-conforming" responses from the data when computing the fit statistics. Table 5.2.10 shows how many such cases were removed from the analysis for each cluster for Series 203.

Table 5.2.10
Rate of Speaking Responses Removed from Fit Analysis S203

Cluster	No. of Responses	No. of Responses Removed	Percent of Responses Removed

1-2	271,634	10,070	3.7%
3-5	254,483	7,371	2.9%
6-8	152,490	5,782	3.8%
9-12	140,899	4,298	3.1%

Removing these items from the analysis helped to lower the outfit value for many of the Speaking tasks. However, there were still some high outfit statistics that were caused by "across folder non-forming" responses. These responses greatly increase the outfit values, especially for the Folder A tasks. Since Folder A tasks were very easy, any deviation would cause large outfit statistics. For the speaking tasks, outfit statistics do not seem to indicate potential problems with the tasks.

Note also that the Kindergarten test used a new format starting with Series 200 (2008-2009). It was equated to Series 103 through a separate study, reported on in MacGregor, Kenyon, Gibson, and Evans (2009). Thus, the column labeled "Anchored?" is not included in Table H for the Kindergarten test.

5.2.11 Complete Raw Score to Scale Score Conversion Table(Table I)

The next table in this section, Table I, presents the raw score to scale score conversion for the test form. The first column shows all possible raw scores. The next one to four columns show the corresponding scale score for each grade level in the cluster. Note that for Listening and Reading items on Tier A, these have been capped to the scale score that represents the proficiency level score of 4.0. On Tier B, these have been capped to the scale score representing the proficiency level score of 5.0.

The next column shows the *conditional* standard error (i.e., from the Rasch analysis) in the metric of the scale score. The last two columns show a lower bound (i.e., the scale score minus one standard error) and an upper bound (i.e., the scale score plus one standard error) around the scale score. In some cases the resulting lower bound is below 100, which has been set as the lowest score on the scale. In those cases, the lower bound has been set at 100.

As can be clearly seen from the table, on any dichotomously-scored test form, standard errors are very large at the lowest and highest ends of the raw score scale. Because of this phenomenon and because the scale scores are combined to form composite scores, the top scale scores for the Listening and Reading forms were often adjusted for an end-of-scale effect on Tier C by allowing the top scale scores to increase only at the same rate as the preceding scale scores. If they were not adjusted, their effect in the composite scores might be excessive.

Thus, if the scale scores towards the high end of the raw score scale were increasing with each raw score by 9 scale points before the group of adjusted scores, then each of the adjusted scores would increase by only 9 scale points each. Because the lower and upper bounds were calculated based on the original logit scores, these adjusted scores do not fall in the middle of the range; they fall toward the lower end of the range, but they always fall ***within*** the range. In other words, the adjusted scale score is a very possible observed score for that number of raw score points obtained.

Because on Tiers A and B the highest possible scores have been capped before the escalation of scale scores due to large standard errors at the highest end of the raw score scale inflates them, there has been no need to make any other adjustment to the scale scores for these tiers at the

extreme high end of the raw score range. Because the point at which scale scores are acapped depends on the proficiency level associated with the score, the caps take effect at lower scores for lower grades within a cluster. In this case the scores have been marked in Table I as capped, and the standard error, and low and high bound for the capped scale score, has been repeated in the final rows of the table.

In addition, at the lower end of the raw score scale, scale scores are truncated when necessary so that the lowest scale score given is the scale score corresponding to a proficiency level score of 1.0. As with the adjusted scores, the standard error and the lower and upper bounds reported in Table I reflect the true scale score, not the truncated score.

5.2.12 Raw Score to Proficiency Level Score Conversion Table (Table J)

The final table, Table J, shows the interpretive proficiency level score associated with each raw score. (Note that in previous Annual Technical Reports some of this information was included in Table I; however, with the grade-level cut scores in effect, we have put this information in a separate table for ease of reading.) The first column in Table J shows the raw score. The remaining columns show the proficiency level score associated with each raw score/scale score for each grade in the cluster, along with the percentage of students in that grade who scored at that raw score/scale score/proficiency level score.

There are two things to note about this table. First, unlike scale scores, which are determined psychometrically and have a one-to-one correspondence to raw scores regardless of the grade level of the student, proficiency level scores are interpretations of the scale score. In Series 100 and 101, cut scores between proficiency levels were determined at the cluster level; thus, for example, in the 3-5 grade-level cluster, a given scale score was associated with the same proficiency level score for students in grades 3, 4, and 5. Such a system, however, fails to take into account that older children can be expected to perform better on the test due to general cognitive growth over and above growth in English language proficiency. This effect can clearly be seen in Tables A and B, where average scores on any test form tend to rise, albeit slightly, by grade level. In order words, we would expect a fifth grader to perform better on the 3-5 grade-level cluster test form than a third grader at the same underlying level of English proficiency. To account for this effect, the WIDA Consortium adopted grade-level cut scores beginning with Series 102 so that, for any given raw score/scale score, the proficiency level score now associated with it differs according to the grade level of the student. (For details on how grade-level cut scores were determined, see Kenyon et al., forthcoming 2013.) The effect of this for Table J is to require a separate column for each grade.

Second, because scale scores are capped on Listening and Reading for Tiers A and B at the scale score corresponding to the proficiency level score of 4.0 (for Tier A) and 5.0 (for Tier B), beginning with Series 102, this capped score is now dependent on the grade level (rather than dependent on the cluster level). These differences in the cap are also shown in Table J on Tiers A and B for Listening and Reading.

For Kindergarten the proficiency level scores are provided based on both the Accountability cut scores and the Instructional cut scores.

6. Analyses of Test Forms: Results

Chapter 6 contains proprietary test information and is not publicly available. State educational agencies (SEAs) may request this information; please contact us at help@wida.us.