World-Class Instructional Design and Assessment

Annual Technical Report for ACCESS for ELLs® English Language Proficiency Test, Series 101, 2005-2006 Administration

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

Prepared by:

Dorry M. Kenyon, Ph.D.
David MacGregor, Ph.D.
Mohammed Louguit, Ph.D.
Bokyung Cho, Ph.D.
Jeong Ran (Willow) Ryu

Center for Applied Linguistics

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®*. Here we just want to note that 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 only 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 by person "n" on item "i"
P_{ni0} = probability of an incorrect response 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 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 of more 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 Listening and Reading multiple choice items on the test forms have mean square fit statistics in the range of .75 and 1.25. Results prove similar for the Speaking tasks. However, for the Writing tasks, it will be noticed that the 30 minute “integrated” Writing tasks tend to overfit the model (i.e., has low infit and outfit mean square statistics). This is a result of the greater weight put on those scores (that is, a weight of 3) compared to the weight on the three shorter (10 minute) tasks, which each carry a weight of 1. Nevertheless, these Writing tasks still fit the range that is "productive for measurement" according to the guidelines above.

5.1.2 Sampling

The data presented in the following tables are based on the full data set of all students administered operational Series 101 of ACCESS for ELLs® in the academic year 2005-2006, with one exception. The raw score to scale score conversion tables (Table I in Section 6) are based on analyses done in the midst of the operational scoring. (For this reason, the item difficulty values in Tables G and H also come from this calibration.) However, analyses to determine the scale score from the raw scores for Listening, Reading and Speaking were based on random samples of at least 500 students per test form.

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, that 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 Section 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 101 represent a partial refreshment of Series 100. That is, while many items were common on both forms, certain folders on Series 101 were replaced with new items. Thus, to place results on Series 101 onto the ACCESS for ELLs® Scale Score, items that were not revised or otherwise changed were anchored to the difficulty values from Series 100, which itself had been anchored to the original the field test. Table E in Section 6 for each test form provides explicit information on the anchor items used for equating Series 101 results to those of Series 100.

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 items on ACCESS for ELLs® was compared by dividing students into two different groupings: first, males versus females; second, students of Hispanic ethnic background versus students of all other backgrounds. (For both analyses, students for whom gender or ethnicity was missing were excluded.) Two commonly used procedures for detecting differential item functioning (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), for these items, the Mantel-Haenszel Chi-square statistic was used, with the M-H common odds ratio that is transferred to the "M-H delta" scale. This procedure compares item level performances of students in the two groups (e.g., males versus females) who are divided into subgroups based on 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 groups were the same across the ability groupings. The M-H common odds ratio is transformed to the "M-H delta" scale to make it symmetrical about zero, where zero has the interpretation of equal odds. If the result is positive, it favors one group; if negative, the other group is favored.

Following guidance proposed by ETS, items were classified 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 were used; however, for many test forms, students scoring some of the lowest and highest raw scores needed 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-stage purification process; that is, items showing DIF in the first stage are removed from the matching variable in the second stage.)

5.1.4.2 Polytomous Items

For these items (on the Writing forms), a similar approach was used based on the Mantel Chi-square statistic and the standardized mean difference following procedures again developed by ETS. The Mantel Chi-square statistic is a conditional mean comparison of ordered response categories (i.e., non-parametric) of the two groups, again conditional on the matching ability variable, which was the total score on the Writing test divided into groups. Based on their total raw score on the Writing test, students were placed into 6

groups. To do so, we determined what the total raw score of a student scoring 1, 2, 3, 4, 5, or 6 in each category 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 since there were so few students in the 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, like 2-1-1, which is a high 1 but not yet a 2); 5-7 category 2; 8-10 category 3; 11-13 category 4; 14-16 category 5; and 17-18 category 6. (The only exception to this was kindergarten Writing tasks, when 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 and calculate the Mantel statistic and determine its probability of significance. This statistic gave an indication of the probability that observed differences were the result of chance but did not indicate how significant that difference was. To indicate how significant the difference was, the standardized mean difference (SMD) between the performances of the two groups being compared was calculated. The standardized mean difference compares the means of the reference and focal 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 is divided by the standard deviation (SD) of the item for the total group. This was also programmed into the Excel spreadsheet.

Following guidance proposed by ETS, items were classified into DIF levels as follows:

- AA (no DIF), when the Mantel Chi-square statistic is not significant; or, when it is significant, 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

When serious DIF is detected on ACCESS for ELLs® items, flagged items are thoroughly investigated for potential causes of DIF and, if there is a clear source, notes are made for refining and clarifying item specifications, the item review process, and/or the external bias review process. If DIF extends across two or more items in a folder and that folder is not scheduled for replacement that year, the folder may be removed from the operational test for the next year.

5.2 Descriptions

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

5.2.1 Raw Score Information (Table A and Figure A)

Table A and Figure 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. In addition, the fourth Writing task (the IT task) is given a weight of 3. Thus, the maximum number of points on each Writing test form varies from 54 for the Tier A forms to 108 for the Tier B and C forms.

For each test form, Table A shows:

- Number of students in the analyses (the number of students who were not absent)
- The minimum observed raw score
- The maximum observed raw score
- The mean (average) raw score
- The standard deviation (std. dev.) of the raw scores

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

5.2.2 Scale Score Information (Table B and Figure B)

Table B and Figure 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, Table B shows:

- Number of students in the analyses
- The minimum observed scale score
- The maximum observed scale score
- The mean (average) scale score
- The standard deviation (std. dev.) of the scale scores

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

Five vertical lines in the figure indicate the five cut scores for the test form; these divide 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 students into all six language proficiency levels. If there are fewer than six sections in Figure B, the first section is **always** language proficiency level 1.)

Note that beginning with Series 101, scale scores for Tier A and Tier B in Listening and Reading have been capped. The highest possible scale score for Tier A is the scale score corresponding to the cut score for language proficiency level 4 (i.e., proficiency level score of 4.0). For Tier B, the highest possible scale score is the score corresponding to the cut score for language proficiency level 5 (i.e., proficiency level score of 5.0). The influence of these caps may be noticed on Table B and Figure B, as well as on many other tables throughout the report.

5.2.3 Language Proficiency Level Information (Table C and Figure C)

Table C and Figure C provide information on the language proficiency level distribution of the students who took the test form based on their performance. Thus, for each test form, each row of Table C shows:

The WIDA language proficiency level designation (1 to 6)

The number of students (count) whose performance on the test form placed them into that language proficiency level in the domain being tested

The percent of students, out of the total number of students taking the form, who were placed into that language proficiency level in the domain being tested

Figure C shows the same information graphically. The horizontal axis shows the six WIDA language proficiency levels. The vertical axis shows the percent of students. Each bar shows the percent of students who were placed into each language proficiency level in the domain being tested based on this test form. (Note that for some domains for kindergarten and Tier A tests, it was not possible to place into all language proficiency levels. Again, if there are fewer than six sections in Figure B, the first section is **always** language proficiency level 1. Table C and Figure C also clearly show the effect of the scoring cap on Tiers A and B.

5.2.4 Scaling Equation Table (Table D)

For each test form, Table D provides the scaling equation for that domain (see 5.1.3 above). This is the equation used to convert an examinee's ability measure into the scale score. Because ACCESS for ELLs® is vertically equated, 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 WIDA 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 101 also appeared on Series 100. 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 form 101 version of that test form) to the previous year's test (i.e., the form 100 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 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 final 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 effects 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 101 test form and on the Series 100 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. A rating scale model was used (see 5.1.1 above) and it was modeled that the same step difficulties would be used for the rating scale at each grade

level cluster and tier. Thus, this information is exactly the same for all Writing tasks. In the case that the difficulty measure was not anchored for any Writing task on a form, these step difficulties values still provide a method of anchoring the results onto a common scale.

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. Again, as in Figure B, five vertical lines indicate the five cut scores 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. If there are fewer than six sections in Figure D, the first section is ***always*** language proficiency level 1.) As would be expected, higher raw scores are required to be placed into higher language proficiency levels. The relative width of each section, 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 B and D, five vertical lines in Figure E indicate the five cut scores 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.) If there are fewer than six sections in Figure E, the first section is ***always*** language proficiency level 1. (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, particularly for the shorter Listening test, 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

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

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 inter-rater reliability is also provided in Table 5. This portion of the table shows, for each of the three or four Writing tasks, the percent 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), nonadj (non-adjacent) and total sum of 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. Otherwise the agreement was counted in the non-adjacent category.

All operational Speaking tests are scored by the test administrator. In this report, information on inter-rater reliability for Speaking provided in Table 5 (except for kindergarten) is based on data from the pilot 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 percent of correct items. The last two columns give information on the Rasch model fit statistics (see above). The first is the average infit mean square statistic; 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 above). The first column gives the DIF level: A, B, or C; AA, BB, or CC for polytomous DIF (i.e., Writing tasks). The next major columns show the contrasting groups in the DIF analyses: either male versus female, or Hispanic versus other ethnicities. Even though DIF may be negligible (category A), 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 equals the sum of three feature scores, which are scored from 1 to 6, for a maximum total of 18. However, papers written in languages other than English or 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 Properties Table (Table H)

Table H presents results of the analyses of all the items or tasks on the test form. The first column provides a complete descriptive name of the item or task. For Listening and Reading items, the first character represents the domain, followed by a number representing a unique number in the item database. The next first two letters indicate the WIDA standard addressed by the item (e.g, MA is the language of mathematics), and the next two characters (e.g., "p3") indicate at what language proficiency level the item is targeted (e.g., Level 3). The next three characters indicate the grade-level cluster. (Note that 91 means 9–12.) The last piece of information indicates the thematic folder name to which the item belongs (e.g., "Lissette").

For the Writing tasks, the naming system is very similar. Note, however, that IT stands for the 'integrated' task; that is, the longer piece of extensive writing that integrates model performance indicators for SI, LA and SS. For Speaking tasks, the naming system is a bit simpler; e.g., S35_AL1. "S" stands for Speaking, "35" shows the grade level cluster, and "A" shows the folder (i.e., the first 3 tasks are in Folder A, addressing the SI performance indicators, the second set of 5 tasks are in Folder B, addressing the LA and SS performance indicators, and the third set of 5 tasks are in Folder C, addressing the MA and SC performance indicators.) The final two characters indicate the language proficiency level (1 to 5) that the tasks in that folder are intended to target.

The next column in Table H presents the item difficulty in logits, while the following 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 following 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 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.

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

The final table in this section, Table I, presents the raw score to scale score conversion table for the test form. The first column shows all possible raw scores. The second shows the corresponding scale score. 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 proficiency level score (e.g., 3.5) corresponding to the scale score. 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.

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 only increase 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 only increase by 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. 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.

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.