BEST Plus Version 2.0 Technical Report

January 2015

CAL
CENTER FOR APPLIED LINGUISTICS

Contents

1. Introduction

This technical report serves as a companion to the *BEST Plus Technical Report: Development of a Computer-Assisted Assessment of Oral Proficiency for Adult English Language Learners*, published in September, 2005. That publication reported on the development of the original version of BEST Plus, originally released in 2003, an updated version of which, BEST Plus Version 2.0, will be made available to the public for the 2015-2016 program year. This report provides technical information related to the operational functioning of BEST Plus and on the development of BEST Plus Version 2.0. Additional information related to the initial development of BEST Plus can be found in the aforementioned *BEST Plus Technical Report* by Center for Applied Linguistics (2005).

1.1. Purpose of BEST Plus

The purpose of BEST Plus is to assess the oral language proficiency of adult English language learners. Oral language proficiency is understood as the underlying competencies that enable the performance of communicative language functions that integrate both listening and speaking skills. BEST Plus assesses ability to understand and use unrehearsed, *conversational*, every-day language within topic areas generally covered in adult English language courses. As an integrative performance assessment of listening and speaking proficiency delivered in a face-to-face mode, the test does not assess an examinee's proficiency in comprehension of *presentational* language of the type encountered, for example, in listening to a radio or television broadcast, or in the production of oral *presentational* language, such as in reading aloud or delivering a prepared speech.

BEST Plus is designed to assess the language proficiency of adult (18 years of age or older) non-native speakers of English who may or may not have received an education in their native language or in English, but who need to use English to function in day-to-day life in the United States. It is designed for the population of adult students typically found in adult education programs in the United States.

1.2. Background to the Test

BEST Plus is a revision of the oral interview section of the original Basic English Skills Test (BEST), which discriminates among the levels of English language proficiency described in the Student Performance Levels (SPL's) (U.S. Department of Health and Human Services, 1998). BEST Plus is also aligned with the requirements of the National Reporting System (NRS) and with the needs of local programs to provide data on learner progress and achievement for comparison across programs and within and across states. These data can also be used for program evaluation and accountability.

A detailed description of the background and development of BEST Plus can be found in the *BEST Plus Technical Report: Development of a Computer-Assisted Assessment of Oral Proficiency for Adult English Language Learners*, published in September, 2005

2. Overview of the Development of BEST Plus Version 2.0

2.1. Refreshment of the item bank

In 2010, CAL conducted a bias and sensitivity review of BEST Plus, the results of which were used to inform the development of new items for BEST Plus Version 2.0. In response to the review, a total of 16 7-item folders (see 2005 *BEST Plus Technical Report* for more details on folder structure) were developed and field tested. These new folders were developed in order to replace folders that had the most concerns from bias and sensitivity reviewers. The new items were seeded with operational BEST Plus items for field testing and a specially-developed algorithm was used to ensure that examinees received both the new items and the old items. The number of students who took each item varied due to the computer-adaptive nature of the test, but the 112 new items were seen by between 6 and 597 examinees, with an average of each item being seen by 194 examinees.

After field testing, statistical analyses were performed on the field test data to ensure that items were functioning adequately. Items that were not delivered to an adequate number of examinees for analysis were not included in the refreshed item bank, and ultimately 90 of the new items developed as part of the whole-folder replacement were included in the item bank for BEST Plus Version 2.0. All of the new field tested items were psychometrically tied to the existing BEST Plus measurement scale. In addition to those 90 items, an additional 21 items from the initial BEST Plus that had been flagged for concerns in the 2010 review were modified, either by replacing a photo or by making construct-irrelevant changes to the question text. Therefore, there were ultimately 111 replaced items in the item bank of 258, for a replacement rate of 43%.

2.2. Revision of print-based locators

In addition to the computer-adaptive version of BEST Plus, the print-based tests (PBTs) of BEST Plus were initially developed following analyses from the computer-administered field test and included three parallel print-based forms (i.e., Forms A, B, and C). Fixed thematic folders of selected items were created based on content considerations and psychometric qualities, particularly item difficulty, to ensure content coverage and an appropriate difficulty challenge for the three level tests (Level 1 - low, Level 2 - mid, and Level 3 - high) included in each form. The goal was to create a testing environment as similar as possible to that of the computer-adaptive version of the test, in which the challenge level of the items asked was most appropriate to the current ability level of the student being tested.

To achieve this goal, the six warm-up questions from the personal identification folder on the computer-adaptive version, along with two additional more challenging questions, served as locator questions across all three forms of the print version. These eight questions are thus the same on all three forms. The total scores on these locator questions are to be used to determine which level (1, 2, or 3) will be administered or whether the test should end after the locator. To ensure comparability of scores between the computer-adaptive and print versions, CAL staff conducted a study in 2010 with 84 students. Eight trained and experienced BEST Plus test administrators conducted the testing. Each student in the study took both the computer-adaptive version and one of the forms of the print version of BEST Plus. The testing was conducted in June, 2010 as part of the normal (end of course) testing cycle. The administration of the tests followed a balanced design in which half of the students took the computer-adaptive version first and the other half took the print version first.

Three paired-sample t-tests were conducted in order to determine if examinee scores were significantly different on one form of the test over another. The results showed that examinees' scores on the print-based Form A were significantly lower than on the CAT, and examinees who took print-based Forms B and C did not have a significant difference between testing instances. Further investigation revealed that the examinees taking print-based Form A had a higher average proficiency level than those who had taken the Form B and Form C, and that examinees may have been placed into a lower form of the test than was necessary to allow them to exhibit the full extent of their proficiency. These analyses of the functioning of the original BEST Plus locator questions suggested that some examinees were being placed into a lower level than they should have been, creating a ceiling effect on their print-based Form A scores.

Consequently, for BEST Plus Version 2.0, a revision of the locator questions was undertaken in order to remedy the problem and to ensure that examinees would be more accurately placed in each print-based form. Because the problem was that the original locator questions seemed to create an artificial capping on print-based Form A scores for higher proficiency examinees, CAL reviewed the two more challenging locator questions: fami.2.PE and hous.2.EL. Item 'fami.2.PE' is a personal expansion question related to the topic domain of family, and item 'hous.2.EL' is an elaboration question related to housing. These more challenging locator questions should allow mid- and high-level students to have an opportunity to demonstrate their higher-level speaking skills, so that they can be placed into the appropriate level. The intention is that examinees' responses to the higher-level questions allows them to take the appropriate level of the PBT so that they do not experience a ceiling effect.

The goal of the revision of the locator questions was to utilize questions that give higher level examinees an opportunity to use more complex language, thus receiving higher scores that will place them into the appropriate level test. Because all BEST Plus items were calibrated on a common scale of Rasch logit in terms of their difficulty levels, a more difficult item is expected to elicit a broader range of speaking ability for mid- and high-performing examinees. The difficulty level of the original challenging locator (fami.2.PE) was 0.21 logits. A new locator of the same item type and topic domain was selected, and it is a more difficult item at 0.87 logits (fami.3.PE). Next, as a result of the fairness and sensitivity review described in section 2.1, the locator question that was identified as hous.2.EL was retired from the BEST Plus item bank for Version 2.0 (the retired item will be referred to hereafter as old.hous.2.EL), so a replacement locator item was sought that would be more difficult. The item that replaced hous.2.EL in the greater item bank (new.hous.2.EL) was examined. The item difficulty of old.hous.2.EL was 1.13, whereas the item difficulty for new.hous.2.EL was 1.46. Therefore, new.hous.2.EL was selected.

As a result of these revisions, the new set of locator questions included the six warm-up questions from the personal identification folder on the computer-adaptive version, item 'new.hous.2.EL', and item 'fami.3.PE'. To confirm that the revision had the desired effect, CAL conducted a number of studies with students from adult ESL programs in several different states, using the BEST Plus Version 2.0 forms of the PBT and the revised locator items that are common across the three forms of the print version.

In the first comparability study, 32 students, representing the full range of levels and class sessions (morning, afternoon, evening), took both the computer-adaptive version and the print-based Form A. The study design called for all students to take both versions of the test on the same day. The administration of the tests followed a balanced design in which half of the students took the computer-adaptive version first and the other half took the print-based Form A first. The testing was conducted in November, 2014.

CAL staff compared the descriptive statistics of the examinees' scale scores on the computer-adaptive version and on the print-based Form A; examined the correlations between scale scores achieved on the two test versions to analyze their relationship; examined the extent to which scale scores differed between the two versions; and examined the degree to which the examinees were placed into the same NRS EFL by both test versions. Table 1 presents the descriptive statistics for the 32 students across the computer-adaptive version and the print version of Form A. It demonstrates the similarity of the average performance across the students in the study.

Table 1. Descriptive statistics for computer-adaptive version and print-based Form A

	N	**Mean**	**Std. Deviation**	**Min**	**Max**
Computer-adaptive Version	32	477.31	89.67	340	801
Print-based Form A	32	474.97	93.22	338	711

Table 2 presents the correlation and scale score difference between the computer-adaptive version and print-based Form A. In this comparability study, the correlation between the scale scores from the two test versions was .890. Considering that BEST Plus is a performance-based assessment, this figure is quite high and provides evidence that examinees will obtain similar results between the computer-adaptive version and the print version of Form A. Additionally, to examine whether there was a statistically significant difference between scale scores on the computer-adaptive version and on the print-based Form A, CAL staff conducted a paired-sample t test. The results, presented in Table 2, show that the difference between performances on the computer-adaptive version and the print version of Form A was not statistically significant, which ensures the comparability of scores between the computer-adaptive version and the print-based Form A of BEST Plus.

Table 2. Correlation and paired difference of scale scores for computer-adaptive version and print-based Form A

	Correlation	**Paired-difference mean**	***t***	***df***	***p***
Computer-adaptive version vs. Print-based Form A	.890	2.344	.307	31	.760

In the second comparability study, CAL staff collected responses from 67 students, representing the full range of levels and class sessions (morning, afternoon, evening). The 67 students were divided into two groups (33 students in one and 34 students in the other). The first group (N=33) took both the computer-adaptive version and the print-based Form B, and The second group (N=34) took both the computer-adaptive version and the print-based Form C. The study design called for each student to take two versions of the test on the same day. Within each group, the administration of the tests followed a balanced design in which half of the students took the computer-adaptive version first and the other half took the print version first. The testing was conducted in December, 2014.

CAL staff compared the descriptive statistics of the examinees' scale scores on the computer-adaptive version and on the print version of Forms B and C, respectively; examined the correlations between scale scores achieved on the two test versions to analyze their relationship; examined the extent to which scale scores differed between the two versions; and examined the degree to which the examinees were placed into the same NRS EFL by both test versions. Table 3 and Table 4 present the descriptive statistics for the 33 students in the first group and 34 students in the second group across the computer-adaptive version and the print version of Forms B and C, respectively. It demonstrates the similarity of mean performance across the students within each group in the study.

Table 3. Descriptive statistics for computer-adaptive version and print-based Form B

	N	Mean	Std. Deviation	Min	Max
Computer-adaptive Version	33	530.79	107.34	336	844
Print-based Form B	33	544.00	85.66	362	742

Table 4. Descriptive statistics for computer-adaptive version and print-based Form C

	N	Mean	Std. Deviation	Min	Max
Computer-adaptive Version	34	526.76	101.48	361	731
Print-based Form C	34	544.38	80.36	390	752

Table 5 presents the correlation and scale score difference between the computer-adaptive version and the print version of Forms B and C, respectively. In this comparability study, the correlation between the scale scores from the computer-adaptive version and the print-based Forms B and C was .882 and .865, respectively. Considering that BEST Plus is a performance-based assessment, these figures are quite high and provide evidence that examinees will obtain similar results between the computer-adaptive version and each of the two print-based Forms B and C. Additionally, to examine whether there was a statistically significant difference between scale scores on the computer-adaptive version and on each of the two print-based forms, CAL staff conducted paired-sample t tests. The results, presented in Table 5 show that the difference between performances on the computer-adaptive version and the print version of Form B was not statistically significant. The difference between performances on the computer-adaptive version and the print version of Form C was also not statistically significant. These results ensure the comparability of scores between the computer-adaptive version and the print-based Forms B and C of BEST Plus.

More importantly, these results compare favorably with the non-significant difference of scale scores between the computer-adaptive version and the print-based Form A. Thus, whether examinees took the computer-adaptive version and any of the three forms of the print version, the difference in performances between the computer-adaptive and print versions, in terms of the final scale scores, was not likely to be statistically significant. These important findings lend support to the use of the revised locator in the placement procedures of the print version of BEST Plus.

Table 5. Correlation and paired difference of scale scores for computer-adaptive version and print-based Forms B and C

	Correlation	**Paired-difference mean**	***t***	***df***	***p***
Computer-adaptive version vs. Print-based Form B	.882	-13.212	-1.477	32	.149
Computer-adaptive version vs. Print-based Form C	.865	-17.618	-1.999	33	.054

3. Reliability

3.1. Classification accuracy and consistency

Research on consistency of NRS classifications based on performances on BEST Plus was conducted using one operational sample drawn from testing sites in different states that agreed to share data with CAL for research purposes. One operational sample (n=1,353) of tests administered between program years 2011–2013 was examined (henceforth referred to as the 2011-2013 operational data set). The sample drew from testing sites in different locations in the United States that have shared their data with CAL for research purposes. Table 6 describes the sample in terms of the total number of examinees as well as the number and percentage of examinees observed in each NRS EFL. Note that the NRS classification for the current analysis was based on the updated cut points described in section 6 of this report. Levels 1 to 6 are the six NRS EFLs, while level 7 represents those who have exited from the NRS EFLs.

Table 6. Frequency distribution of the NRS educational functioning levels for one operational sample

	Operational Sample N=1,353	
NRS level	Frequency	Percent
1	245	18%
2	401	30%
3	155	11%
4	189	14%
5	146	11%
6	89	7%
7	128	9%
Total	**1,353**	**100%**

Data collected from the examinees were analyzed using the methods outlined and implemented in Livingston and Lewis (1995) and Young and Yoon (1998). This methodology seeks to estimate classification accuracy and consistency of data from a single administration of a test. Using a four-parameter beta distribution, the approach seeks to model the underlying distribution of the true scores to examine accuracy and also uses the hypothetical distribution of one of many possible parallel forms to determine consistency. Given the modeled distributions, the reliability of the test, and the location of the cut scores vis-à-vis the modeled score distributions, overall accuracy and consistency indices are produced by comparing the percentage of students classified across all categories the same way by the observed distribution against the modeled true score distribution and against a (hypothetical) parallel form distribution. Based on the statistical modeling, these indices indicate the percent of all students who would be classified into the same language proficiency level by both the administered test and either the true score distribution (accuracy) or a parallel test (consistency). The analysis also provides an estimate of Cohen's kappa statistic, which is a very conservative estimate of the overall classification since it corrects for chance.

In addition to overall accuracy and consistency, the analysis also provides accuracy and consistency conditional on the language proficiency level. These indices examine the percent of students classified by both tests into a level divided by all students classified into that level according either to the true score distribution (accuracy) or based on a parallel test (consistency).

The analysis also provides what may be the most important set of indices, those at the cut points. At every cut point, using the modeled true score distribution (accuracy), the analysis provides the percentage of students who are consistently placed above and below the cut score, as well as those who are false positives and false negatives. For consistency, only the percent of students classified consistently above and below the cut score is calculated. Thus, for example, to evaluate the degree of confidence one can have in a decision made based on the BEST Plus score as to whether students are being accurately classified into level 3 (NRS EFL High Beginning ESL) or not, one can look at the accuracy index provided in the table for the cut point labelled 2/3; i.e., between level 2 (NRS EFL Low Beginning ESL) and level 3 (NRS EFL High Beginning ESL).

Table 7 presents the results of analysis of the scores from the sample of 1,353 examinees from the operational data. The first block of two rows in the table shows the overall indices in terms of accuracy, consistency, and kappa. The second block of rows shows the accuracy and consistency conditional on NRS EFLs. Finally, the last block of rows shows the accuracy and consistency at cut points.

Table 7. Accuracy and Consistency of Classification indices: NRS EFL (n=1,353)

Overall Indices	Accuracy	Consistency		Kappa (k)	
	0.557	0.460		0.346	
Conditional on Level	Level	Accuracy		Consistency	
	1	0.748		0.629	
	2	0.652		0.546	
	3	0.282		0.217	
	4	0.383		0.292	
	5	0.387		0.284	
	6	0.385		0.269	
	7	0.821		0.687	
Indices at Cut Points		Accuracy			
	Cut Point	Accuracy	False Positives	False Negatives	Consistency
	1/2	0.915	0.048	0.038	0.876
	2/3	0.873	0.066	0.062	0.826
	3/4	0.877	0.065	0.058	0.833
	4/5	0.903	0.046	0.051	0.863
	5/6	0.938	0.033	0.029	0.908
	6/7	0.961	0.024	0.015	0.942

In the first block in Table 7, the overall indices indicate the percent of all examinees who would be classified into the same NRS EFL by both the administered test (observed data) and either the modeled true score distribution (accuracy) or a modeled parallel test (consistency). The indices range from 0 to 1; the closer they approach to 1, the higher the consistency of agreement between the two classifications being compared. Table 7 shows that the overall estimate for accuracy (that is, being placed into the same NRS EFL by both the observed scores and the model true scores) is .557, while the overall estimate for consistency (that is, being placed into the same NRS EFL by both the observed scores and modeled scores from one of many potential parallel forms) is .460. The block also contains Cohen's kappa, a very conservative estimate of the overall classification because it corrects for chance. Cohen's kappa statistic is .346.

The second block in Table 7, labeled "conditional on level," presents accuracy and consistency conditional on each NRS EFL (i.e., the percentage of examinees classified by both tests into a level). The accuracy of classification conditional on NRS EFLs ranges from .282 to .821, and the consistency of classification conditional on NRS EFLs varies between .217 and .687.

The third block, labeled "indices at cut points," looks at every cut point and, using the true score distribution (accuracy), shows the percentage of examinees who were consistently placed above and below that cut point, as well as those who were false positives and false negatives. The consistency column shows the percentage of examinees classified consistently above and below the cut point using the modeled parallel test. Thus, for example, to evaluate the degree of confidence one can have in a decision as to whether examinees are being accurately classified into NRS EFL High Intermediate ESL (Level 5), one can look at the accuracy and consistency indices provided in the above table for the cut point between NRS EFL Low Intermediate ESL (Level 4) and NRS EFL High Intermediate ESL (Level 5), i.e., at the row labeled 4/5, which is .903 for accuracy and .863 for consistency. The False Positive and False Negative indices suggest that there is low probability that an examinee may be incorrectly classified to a level that is higher (Positive) or lower (Negative) than his or her true ability. The lowest accuracy of NRS EFL classification at any cut points is .873, and the lowest consistency of NRS EFL is .826.

It should be emphasized that classification accuracy and consistency at the cut points (i.e., the third block, labeled "indices at cut points") are perhaps the most important of all when using any of these indices as an absolute criterion in making decisions as to which students have reached a particular NRS EFL. For BEST Plus test takers, the scale scores are used for making decisions about student placement into adult ESL programs. Overall, the accuracy and consistency of NRS classification at each cut point is very high, confirming that BEST Plus is classifying examinees accurately and consistently with respect to the NRS EFLs.

3.2. Interrater reliability

In November 2014, CAL conducted an inter-rater reliability study on the computer-adaptive version and print versions of BEST Plus, using the updated item bank prepared for Version 2.0. The purpose of this study was to ensure a high level of consistency across test administrations and test delivery modes, regardless of the test administrator. Inter-rater reliability is an important area to study for performance assessments that require the use of a holistic scoring rubric to rate examinee performance.

The study involved 49 students from one adult ESL program, drawn from all program levels, and two teams of raters: Team A and Team B. Each team consisted of one program test administrator, one experienced CAL test administrator, and one novice CAL administrator who had received about three hours of training in scoring BEST Plus. In each team, the experienced CAL administrator conducted all of the tests, observed by the other two team members.

The tests used were the computer-adaptive version and a print version consisting of a subset of BEST Plus items from the updated item bank. Each student was tested twice: once using the computer-adaptive version by the Team A test administrators, and once using the alternate print version by the Team B test administrators. Within each test session, while the experienced CAL test administrator administered the test, the other two team members sat beside the examinee and observed the administration. In the computer-adaptive group, the test administrator scored student responses on the computer; the other two members recorded their scores on specially-designed score sheets. In the print-based group, the test administrator scored student responses on the print-based test sheet; the other two members recorded their scores on specially-designed score sheets.

Analyses were first conducted on each examinee's total raw scores across all of the test items administered, since these total scores determine how the computer-adaptive program estimates the examinee's ability. This analysis looked at the correlations between each pair of raters within each team, as well as the average correlation across the three pairs of raters. Next, an additional analysis was conducted, using BEST Plus scale scores from the print version, to examine correlations between rater pairs. This analysis was conducted using scale scores (rather than raw scores) because the scale scores provide substantive interpretations in terms of the NRS educational functioning levels.

Table 8 presents the descriptive statistics for the three raters in Team A (computer-adaptive version). The table shows that for Team A, the mean scores across all students and all items for total scores and for all three subscales are very close. They vary least for the Listening Comprehension subscale. For Language Complexity and Communication, there is little variation between the CAL experienced and the CAL novice raters; the program administrator is slightly higher on the Communication subscale and slightly lower on the Language Complexity subscale, but still very close to the others. These data indicate that across the examinees, the raters were applying the scoring rubric with a great deal of consistency in the computer-adaptive delivery mode.

Table 8. Descriptive Statistics for Team A: Computer-adaptive Version

		N	Min	Max	Mean	Std. Dev.
TOTAL SCORE	Program_Admin-A	49	41	177	118.12	32.417
	CAL_Novice-A	49	44	168	117.24	31.071
	CAL_Experienced-A	49	51	168	117.29	29.368
Listening Comprehension	Program_Admin-A	49	13	50	37.04	9.060
	CAL_Novice-A	49	15	50	37.35	8.722
	CAL_Experienced-A	49	16	50	37.18	8.555
Language Complexity	Program_Admin-A	49	7	52	24.65	10.197
	CAL_Novice-A	49	7	49	25.12	9.957
	CAL_Experienced-A	49	10	47	25.06	9.306
Communication	Program_Admin-A	49	15	75	56.43	15.262
	CAL_Novice-A	49	22	75	54.78	14.185
	CAL_Experienced-A	49	25	73	55.04	12.961

Table 9 presents correlations between pairs of scores awarded by different rater pairs within Team A. The first column shows the scoring category. The next three columns show the correlation of the scores of each rating pair. The final column shows the simple average of the three paired correlations. The correlations between raters in each pair within Team A on the total score and all subscores were extremely high, ranging from a high of .98 between all three rater pairs on the Listening Comprehension subscale and between the CAL-experienced/program-administrator pair and the CAL-experienced/CAL-novice pair on the total score to a low of .93 between the program administrator and the CAL experienced administrator and between the program administrator and the CAL novice administrator on the Language Complexity subscale. The average rater-pair correlations on the total score and all sub-scores are extremely high, ranging from a high of .98 on the total score and the Listening Comprehension subscale to a low of .94 on the Language Complexity subscale.

Table 9. Pearson Correlations for Team A: Computer-adaptive Version

Category	Rating Pair			Average
	Prog/CAL-Exp	Prog/Cal-Nov	CAL-Exp/Cal-Nov	
TOTAL SCORE	.98	.97	.98	.98
Listening Comprehension	.98	.98	.98	.98
Language Complexity	.93	.93	.97	.94
Communication	.96	.95	.97	.96

Table 10 presents the descriptive statistics for the three raters in Team B (print-based BEST Plus). The table shows that for Team B, the mean scores across all students and all items for total score and for all three subscales are close. Although the CAL experienced rater was higher on the Language Complexity subscale by 2.4 points than the mean of the other two raters and was lower on the Communication subscale by a bit more than 1 point than the mean of the other two raters, these variations are not extreme given the size of the standard deviations. As with Team A, these data indicate that the raters in Team B were consistently applying the scoring rubric in the print-based delivery mode. It should be noted that raw scores from Table 8 Table 10 and cannot be compared directly because the test items administered were not the same between the computer-adaptive and the print versions and the number of items the students received between the two versions differed. For test versions with different items and length, only scale scores, to which raw scores are transformed onto a common scale, can be directly compared.

Table 10. Descriptive Statistics for Team B: Print Version

		N	**Min**	**Max**	**Mean**	**Std. Dev.**
TOTAL SCORE	Program_Admin-B	49	18	145	101.69	29.031
	CAL_Novice-B	49	15	140	102.45	26.271
	CAL_Experienced-B	49	15	149	103.47	26.950
Listening Comprehension	Program_Admin-B	49	6	42	32.31	8.581
	CAL_Novice-B	49	5	42	32.55	7.743
	CAL_Experienced-B	49	5	42	32.67	7.358
Language Complexity	Program_Admin-B	49	3	45	21.20	8.490
	CAL_Novice-B	49	3	38	20.78	7.145
	CAL_Experienced-B	49	3	48	23.39	9.453
Communication	Program_Admin-B	49	9	63	48.18	13.458
	CAL_Novice-B	49	7	63	49.12	12.397
	CAL_Experienced-B	49	7	63	47.41	11.740

Table 11 presents correlations between pairs of scores awarded by different rater pairs within Team B. While somewhat lower than the correlation in Table 9 (which may be expected given the greater variations in average scores), the correlations between raters in each pair in Team B remain quite high, averaging .96 for the total score and Listening Comprehension subscale. The highest observed correlation, .98, was between the CAL experienced and novice raters on the total score and the Listening Comprehension subscale. The lowest correlation (.85) was between the CAL experienced rater and the program administrator on the Language Complexity subscale.

Table 11. Pearson Correlations for Team B: Print Version

Category	Rating Pair			Average
	Prog/CAL-Exp	Prog/Cal-Nov	CAL-Exp/Cal-Nov	
TOTAL SCORE	.94	.95	.98	.96
Listening Comprehension	.95	.96	.98	.96
Language Complexity	.85	.87	.96	.89
Communication	.94	.94	.96	.95

This examination of the data from the two groups of raters shows that the degree of inter-rater reliability that can be achieved using the BEST Plus scoring rubric is quite high, even including novice raters. The average correlation for the total score for both teams of raters was very high: .98 for Team A and .96 for Team B. For Listening Comprehension, it was also .98 for Team A and .96 for Team B. For Language Complexity, it was .94 for Team A and .89 for Team B. For Communication, it was .96 for Team A and .95 for Team B.

CAL staff further conducted a study to examine inter-rater reliability using BEST Plus scale scores from the print version, since these scale scores are linked to students' level of proficiency in terms of NRS educational functioning levels. Only scale scores from the print version were available for the current study because the print-based scoring system allows CAL staff to enter ratings from all three raters into the score management software, after the test has taken place, to arrive at the final scale scores. (This was not possible for the computer-adaptive scoring system since it can only produce scale scores by the single rater who administered the computer-adaptive test.) The analysis looked at correlations between each pair of raters in Team B, as well as the average correlation across the three pairs of raters.

Table 12 presents the descriptive statistics of BEST Plus scale scores from the three raters in Team B. The table shows the mean scale scores across all students do not vary much. There is a range of only about 7 scale points between the lowest mean (508.88 – the CAL novice rater) and the highest mean (515.94 – the CAL experienced rater), with the CAL novice rater and the program administrator only slightly more than 1 scale point apart. This variation is not extreme given the size of the standard deviations and given that the standard error of measurement on BEST Plus scale score is 20 points on a scale from 88 to 999. As with the results using total raw scores, these data indicate that raters can apply the scoring rubric with a high degree of consistency in arriving at students' scale scores.

Table 12. Descriptive Statistics for BEST Plus Scale Scores: Print Version

	N	Min	Max	Mean	Std. Dev.
Program_Admin	49	302	725	510.12	90.226
CAL_Novice	49	302	685	508.88	77.835
CAL_Experienced	49	302	769	515.94	88.716

Table 13 presents the correlation between pairs of BEST Plus scale scores awarded by different rater pairs within Team B. Results indicate that the correlations between raters in each pair are quite high, averaging .94 across the three rater pairs. The highest observed correlation, .97, was between the CAL experienced and novice raters. The lowest correlation, .93, was observed between the CAL experienced rater and the program administrator and between the CAL novice rater and the program administrator.

Table 13. Pearson Correlations for BEST Plus Scale Scores: Print Version

Rating Pair			**Average**
Prog/CAL-Exp	**Prog/Cal-Nov**	**CAL-Exp/Cal-Nov**	
.93	.93	.97	.94

This examination of BEST Plus scale scores from the rater pairs for the print version provides additional evidence that the degree of inter-rater reliability that can be achieved using the BEST Plus scoring rubric is quite high, even when including a novice rater.

4. Validity

4.1. Content alignment to NRS

On August 14 and 15, 2014, CAL hosted a content alignment study at its Washington, DC offices with five experts in adult ESL who served as reviewers. The purpose of the study was to align the BEST Plus Version 2.0 items to the National Reporting System (NRS) Educational Functioning Levels (EFLs) and to ensure that all aspects of the descriptors of the NRS EFLs are addressed by the test items. The findings and evidence for content validity in this section are the result of that study.

4.2. Types and levels of skills in test items and NRS educational functioning levels

4.2.i. Mapping of items to NRS EFLs

Each BEST Plus item was mapped to one NRS EFL based on the judgments of the reviewers. The final NRS EFL to which each item was assigned was determined based on the NRS EFL that was assigned by the greatest number of reviewers. For those items that were reviewed during the August session, after discussion, every item fell on either one or two levels, in which case the level assigned by three or more reviewers was taken. Because items that were reviewed in October using the online survey did not have an opportunity for discussion, any of these that fell across more than two levels were assigned to the level falling in the middle of the range. Table 14 summarizes the number of items at each level.

Table 14. BEST Plus items by NRS EFL

NRS EFL	Items	Percentage
Beginning ESL Literacy	41	16%
Low Beginning ESL	47	18%
High Beginning ESL	64	25%
Low Intermediate ESL	34	13%
High Intermediate ESL	46	18%
Advanced ESL	26	10%
Total	**258**	**100%**

Table 15 shows a detailed mapping of each item in the BEST Plus 2.0 item bank to an NRS EFL. This table shows all of the content folders of BEST Plus, each of which contains seven items, with the item types listed in the left-hand column. Because the warm-up items do not follow the same format as the other folders, they are presented separately in Table 16.

Table 15. Alignment of BEST Plus 2.0 items to NRS EFLs

Gen. Domains	Personal																Occupational						Public																	
Specific Domains	Person-al ID	Health				Family/ Parenting			Consumer-ism			Housing			Recreation/ Entertain-ment			Getting a job			On the job			Civics			Community Services			Transporta-tion/Direc-tions			Education			Weath-er	
Item Folder #	1	1	2	3	4	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2
Photo Description	For this folder see note below	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1
Entry Question		3	2	3	3	3	2	2	2	3	3	3	2	2	3	3	3	4	3	3	2	3	3	3	3	3	3	4	3	3	3	3	3	2	3	2	4			
Yes/No Question		2	2	1	2	2	3	2	2	2	2	2	3	2	2	2	1	2	2	3	2	1	2	2	3	3	3	2	3	2	2	2	2	1	2	2	2			
Choice Question		3	3	3	3	3	3	2	3	3	3	3	3	2	2	2	2	3	3	3	3	2	3	3	3	3	3	3	3	2	2	3	3	2	3	3	2			
Personal Expansion		4	4	4	5	5	4	4	3	4	4	4	5	3	3	4	4	4	4	5	3	3	5	4	5	4	4	5	4	4	5	4	4	4	4	4	4			
General Expansion		5	4	5	5	5	5	5	4	5	5	5	4	5	4	5	5	5	4	6	5	5	5	5	5	4	5	5	5	5	5	5	5	4	4	5	5			
Elaboration		6	5	6	6	6	6	5	5	5	6	6	6	5	5	6	6	5	5	6	6	5	6	6	6	5	6	6	5	6	6	6	6	6	6	6	6			

Table 16. Alignment of Warm-up Folders to BEST Plus

Personal ID Folder 1	
Warm-up 1	1
Warm-up 2	2
Warm-up 3	2
Warm-up 4	3
Warm-up 5	1
Warm-up 6	3

Table 15 & Table 16 Key	
1	Beginning ESL Literacy
2	Low Beginning ESL
3	High Beginning ESL
4	Low Intermediate ESL
5	High Intermediate ESL
6	Advanced ESL

4.2.ii. Relationship of items to levels and skills used to describe NRS EFLs

The six NRS EFLs were broken down into 28 component descriptor statements. Reviewers were then asked to assign one or more descriptor statements to each item in the BEST Plus Version 2.0 item bank. It was determined that an item covered a given descriptor if more than half (i.e., three or more) of the five reviewers had selected that level for that item (Chi et al., 2011). Table 17 shows these statements, the NRS EFL from which they come, and the number of items in the BEST Plus item bank that are represented by that descriptor.

Table 17. Count of BEST Plus 2.0 items per descriptor

NRS EFL	Descriptor	Number of items
Beginning ESL Literacy	Individual cannot speak or understand English.	0
	Individual understands isolated words or phrases.	40
Low Beginning ESL	Individual can understand basic greetings, simple phrases and commands.	31
	Individual can understand simple questions related to personal information, spoken slowly and with repetition.	37
	Individual can understand a limited number of words related to immediate needs.	10
	Individual can respond with simple learned phrases to some common questions related to routine survival situations.	22
	Individual speaks slowly and with difficulty.	32
	Individual demonstrates little or no control over grammar.	28
High Beginning ESL	Individual can understand common words, simple phrases, and sentences containing familiar vocabulary, spoken slowly with some repetition.	52
	Individual can respond to simple questions about personal everyday activities.	45
	Individual can express immediate needs, using simple learned phrases or short sentences.	2
	Individual shows limited control of grammar.	35
Low Interme-diate ESL	Individual can understand simple learned phrases and limited new phrases containing familiar vocabulary spoken slowly with frequent repetition.	26
	Individual can ask and respond to questions using simple learned phrases and limited new phrases containing familiar vocabulary.	23
	Individual can express basic survival needs and participate in some routine social conversations although with some difficulty.	12
	Individual has some control of basic grammar.	19
High Interme-diate ESL	Individual can understand learned phrases and short new phrases containing familiar vocabulary spoken slowly and with some repetition.	42
	Individual can communicate basic survival needs with some help.	3

NRS EFL	Descriptor	Number of items
	Individual can participate in conversation in limited social situations and use new phrases with hesitation.	41
	Individual relies on description and concrete terms.	33
	Individual has inconsistent control of more complex grammar.	26
Advanced ESL	Individual can understand and communicate in a variety of contexts related to daily life and work.	23
	Individual can understand and participate in conversation on a variety of everyday subjects, including some unfamiliar vocabulary, but may need repetition or rewording.	23
	Individual can clarify own or others' meaning by rewording.	13
	Individual can understand the main points of simple discussions and informational communication in familiar contexts.	24
	Individual shows some ability to go beyond learned patterns and construct new sentences.	18
	Individual shows control of basic grammar but has difficulty using more complex structures.	17
	Individual has some basic fluency of speech.	20

Reviewers assigned all NRS EFL descriptors to BEST Plus items, with the exception of the Beginning ESL Literacy descriptor "Individual cannot speak or understand English." The goal of BEST Plus items is to elicit language; therefore, it is not the target of BEST Plus items to elicit an absence of language. Within levels, some descriptors are associated with noticeably fewer items. The High Beginning ESL descriptor, "Individual can express immediate needs, using simple learned phrases or short sentences," was selected for two items. This statement targets an individual's expression of needs, but BEST Plus does not simulate scenarios in which examinees are expressing immediate needs.

At the High Intermediate level, reviewers selected "Individual can communicate basic survival needs with some help" for three items. The concept of being helped by the interlocutor contradicts the necessities of standardized testing; on BEST Plus, the only form of help that can be offered to the examinee is a single repetition of the test question. Negotiation of meaning by the interlocutor/administrator is not possible in order to maintain the standardization of BEST Plus as it is currently designed.

4.2.iii. Effect of revisions to NRS EFLs on content validity

In 2006, the NRS EFLs were revised so that the Beginning ESL level was divided into Low Beginning ESL and High Beginning ESL and the Low Advanced ESL and High Advanced ESL levels were collapsed into Advanced ESL. Although BEST Plus was originally released in 2003, the results of the 2014 content alignment study, using the post-2006 revised NRS EFLs, shows that these revisions to the NRS functioning levels have not had a negative impact on the content validity of BEST Plus. Original development of BEST Plus and of the item specifications was done at a time when the adult education field was using both the original NRS functioning level descriptors and the Student Performance Levels (SPLs). When it was originally developed, BEST Plus was intended for the same learner population for which it currently seeks approval, and items were developed with those learners in mind as well as the prevailing proficiency scales.

Reviewers indicated that there was coverage of aspects of the NRS EFLs from the two new Beginning ESL levels as shown in Table 17 above. The proportions of items assigned to these levels were similar to the other levels, indicating that there is an adequate number of items to cover these two levels.

Results of the content alignment study suggest that items that would previously have been aligned to the High Advanced ESL level are now subsumed by the Advanced ESL level, without having a negative impact on the distribution of items among the NRS EFLs. The number of items falling into the Advanced ESL level is not disproportionately high, and reviewer comments did not suggest that the items that were categorized as Advanced ESL exceeded the difficulty level that would be appropriate at the Advanced ESL level. These responses suggest that the item bank's alignment to the current NRS EFLs has not been negatively affected by the elimination of the High Advanced ESL level.

4.3. Measured skills not associated with the NRS educational functioning levels

Reviewers had an option for each item to indicate if performance was affected by skills beyond the scope of the NRS EFLs. While reviewers made use of this open-ended section, they did not indicate in any cases that successful completion of the item was dependent on skills outside of the scope of the NRS EFLs. In most cases the reviewers made comments related to their justifications for selecting given NRS EFL descriptors or NRS EFLs.

4.4. Aspects of NRS educational functioning levels not covered

The results of the content alignment study indicate that all aspects of each NRS EFL are covered by BEST Plus items, with the exception of "Individual cannot speak or understand English."

4.5. Procedures for establishing content validity

On August 14 and 15, 2014, CAL hosted a content alignment study at its Washington, DC offices with five experts in adult ESL who served as reviewers. The purpose of the study was to align the BEST Plus Version 2.0 items to the NRS EFLs and to ensure that all aspects of the descriptors of the NRS EFLs are addressed by the test items. The session was facilitated by a CAL testing specialist with support from a psychometrician and an additional notetaker.

The study began with an overview of content alignment study and its goals, as well as a discussion of each NRS EFL and its criteria. After reviewing the NRS EFLs, the procedures for the study were discussed and the participants did a practice review of one folder of seven items. After debriefing on this practice set and having their questions answered, the participants were instructed to begin reviewing items.

A total of 258 items were reviewed; this number included all warm-up items and all items projected to comprise the item bank of BEST Plus Version 2.0. Items were reviewed in 10 sets of 14-27 items and reviewers were instructed to analyze one set at a time by reading items in printed booklets and entering their judgments in a corresponding online survey. While an online survey was used to facilitate the data collection and analysis, these reviews took place in person, not remotely. The online survey asked reviewers to indicate for each item the lowest NRS EFL at which an examinee would be able to sufficiently accomplish the task required by the item.

Reviewers were asked to focus on the lowest NRS EFL on the assumption that higher levels subsume the lower levels. The reviewers were also provided with guidelines regarding what it meant to sufficiently accomplish each BEST Plus item type; these guidelines were presented in the printed booklets beside each item. In addition to the overall NRS EFLs, the survey presented the descriptors for each NRS EFL, broken down into individual statements.

Figure 1 shows an example of how these descriptors were presented for one level (High Beginning ESL).

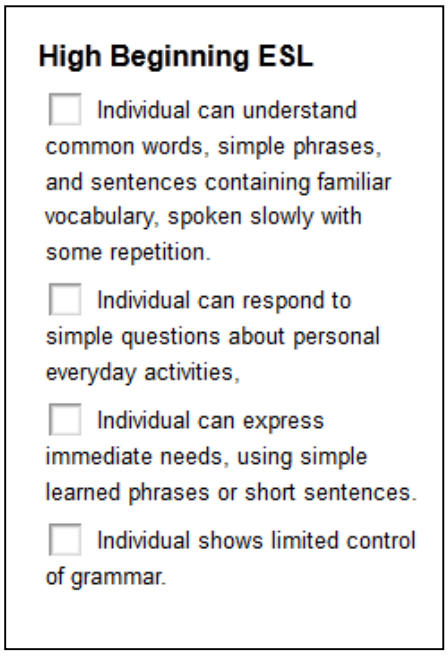

Figure 1 . Example of NRS EFL aspects

For each item, reviewers were instructed to select all of the aspects of the appropriate NRS EFL that were addressed by that item; multiple aspects could be selected. The online survey also provided reviewers with the opportunity to indicate if the item targeted skills outside the scope of an NRS EFL and to provide open-ended comments to explain this response.

After each set of items was reviewed, there was a small break for the psychometrician to review the results from that set. Any items for which reviewer responses spread across more than two NRS EFLs were brought back to the group for discussion. After discussion, reviewers could decide to change the NRS EFL they assigned to that item. A notetaker took notes during the discussion about reviewer concerns and justifications for decisions and also documented the reviewers' new level assignments if they changed. After the discussion rounds, all item judgments fell on either one or two NRS EFLs.

(Note: At the time of the August content alignment study, analysis of the results of new items to be included in BEST Plus Version 2.0 was ongoing. Some items were selected for inclusion in BEST Plus Version 2.0 after the content alignment study was completed. In order to ensure that the content alignment analysis reflected the final BEST Plus 2.0 item bank as closely as possible, a supplemental survey was sent out in October 2014 asking the five reviewers to provide judgments about an additional 17 items not covered in the earlier review. All results presented here represent the content representation of the final BEST Plus Version 2.0 item bank, incorporating results from the August, 2014 content alignment study and the updated information from October, 2014)

4.6. Agreement of subject matter experts' judgments

The intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979) was used in the current study to indicate the degree of consistency among the panel of reviewers in their NRS ratings. The unit of analysis is at the item level, where each item was assigned an NRS EFL by each of the five reviewers. The measurement model is the random-effect model ICC(2,k), where k refers to the number of reviewers. The random-effect model treats the reviewers as random and therefore the generalizability is to a population of similarly qualified and trained reviewers. The ICC was calculated for the NRS EFL assignments given to the items before discussion. For the 17 items that were reviewed in the October follow-up study, there was no discussion session, therefore, there was only one instance of assignment to levels. The ICC among the five reviewers was 0.96. This is a very high ICC, suggesting that the reviewers were very consistent in assigning NRS levels to the items.

4.7. Relationship to instructional hours

Following a complete test administration of the BEST Plus, a score report is generated that expresses the test results in terms of a BEST Plus scale score and an NRS Educational Functioning Level (EFL). Since the launch of the BEST Plus in 2003, a question of great interest among its users has been how many hours of instruction are required for learners to show a gain on the test in terms of an NRS EFL. In 2007, CAL first conducted research to provide empirically-based guidance on this question (Young, 2007). The 2007 paper's descriptive analyses indicated a general trend that the greater the number of hours of instruction that a learner received, the more likely it was that the learner would show gains on the BEST Plus. Young's conclusion included a recommendation that learners be re-tested with the BEST Plus after a minimum of 80 hours of instruction. However, because users of the test have reported that noticeable gains in proficiency seem to be evident in fewer than 80 hours of instruction, and because the Young paper's smallest category of hours of instruction was 60 or fewer hours, it was decided to revisit this question with new test-taker data and to investigate hours of instruction more discretely (0–19 hours, 20–39 hours, etc.). In addition, the cut scores for BEST Plus are now updated (see section 6) following a vigorous standard-setting that related performance on BEST Plus directly to the NRS EFL descriptors. In other words, results from Young (2007) are no longer applicable. Finally, it was decided to refine the methodology used by investigating both meaningful gains in scores and NRS EFL gains. Because of the nature of cut scores, it is possible to show a level gain with only a modest improvement in test scores. Conversely, it is possible to make large improvements within a score band without crossing into the next NRS EFL. It was therefore deemed important to be able to capture and discuss the number of learners that fell into both categories. As a result, the two research questions that are addressed in this section are: (a) what is the relationship between instructional hours and NRS EFL gain on BEST Plus? and (b) what is the relationship between instructional hours and meaningful score gain on BEST Plus?

To attempt to answer these questions BEST Plus pre-test and post-test scores were collected during the 2012–2013 program year from 4,703 adult ESL learners in Minnesota, South Carolina, Utah, and Virginia. The analyses described below included the 2,960 learners for whom there were complete data (pre-test score, post-test score, and instructional hours attended). It is important to note that for these analyses, instructional hours were calculated as hours of actual attendance and not as the total number of hours of the program in which the learner was enrolled (i.e., if a learner was enrolled in a 20-day program for two hours of instruction a day but missed four days, his total instructional hours were 32 and not 40).

As described in detail in section 6, a standard-setting study was carried out to establish new cut scores for the BEST Plus in relation to the NRS EFLs. These cut scores will be in effect from 2015 onwards and were the cut scores used for the current analyses. They are reported below in Table 18.

Table 18. Alignment of NRS EFLs and BEST Plus Scale Scores

NRS EFL	BEST Plus score range
Beginning ESL Literacy	361 and below
Low Beginning ESL	362 to 427
High Beginning ESL	428 to 452
Low Intermediate ESL	453 to 484
High Intermediate ESL	485 to 524
Advanced ESL	525 to 564
Exit from NRS EFLs	565 and higher

As mentioned above, 2,960 learners were included in the final data set. Table 19 shows the numbers and percentages of learners per their initial NRS EFL. Across the first five levels there is a moderately even distribution of learners, with the percentage of learners in each of the first five levels ranging from 12% to 32%. Although the majority of the learners were at the two lowest NRS EFLs (58%), a substantial proportion (37%) entered their classes at the third through fifth level. Less than 5% of the learners scored at the highest level (Advanced ESL) or scored out of the NRS Educational Functioning Levels on their pre-test.

Table 19. Distribution of Learners in NRS EFLs based on BEST Plus Pre-test

Initial NRS EFL	Number	Percentage
Beginning ESL Literacy	761	26%
Low Beginning ESL	940	32%
High Beginning ESL	368	12%
Low Intermediate ESL	397	13%
High Intermediate ESL	362	12%
Advanced ESL	117	4%
Exit from NRS EFLs	15	1%
Total	**2960**	**100%**

Table 20 shows the percentage of learners from each initial NRS EFL who demonstrated at least one NRS EFL gain based on their post-test scores after having received instruction. Table 19 above showed that the majority (58%) of the learners began at the two lowest levels, but as Table 20 shows, a majority of learners at all levels, from 63% to 83%, showed a gain in NRS EFL and that 77% overall showed a gain of one or more levels.

Table 20. Number of Learners Attaining NRS EFL Gain by NRS EFL

Initial NRS EFL	Learners who attained level gain
Beginning ESL Literacy n=761	635 (83%)
Low Beginning ESL n=940	672 (71%)
High Beginning ESL n=368	311 (85%)
Low Intermediate ESL n=397	314 (79%)
High Intermediate ESL n=362	272 (75%)
Advanced ESL n=117	74 (63%)
Total n=2960	**2278 (77%)**

It is useful to see that learners are making level gains after receiving instruction, but a question of major interest is how many hours of instruction were received by the learners who made level gains, as the amount of instruction received by the learners in this study varied. To investigate this question, hours of instruction were divided into 20-hour increments, from 0–19 to 120–139, with the final category being 140 or more hours. Table 21 shows the number of learners (in column 1) within each 20-hour category and the number and percentage of those learners who showed an NRS EFL gain (in the second column).

Table 21. Number of Learners Attaining NRS EFL Gain by Instructional Hours

Instructional hours	Learners who attained level gain
0 to 19 Hours n=41	15 (37%)
20 to 39 Hours n=160	103 (64%)
40 to 59 Hours n=320	246 (77%)
60 to 79 Hours n=1392	1075 (77%)
80 to 99 Hours n=460	367 (80%)
100 to 119 Hours n=225	174 (77%)
120 to 139 Hours n=126	105 (83%)
140 or More Hours n=236	193 (82%)
Total n=2960	**2278 (77%)**

What is interesting about this table is that from the second category of hours and upwards, a majority of the learners showed at least a level gain. That is, while only a little more than a third of the learners in the initial category of 0–19 hours showed a level gain, 64% of the learners in the 20–39 hour category showed level gains. This increases to 77% for both the 40–59 and 60–79 hour categories and then goes up to 80% for the 80–99 hour category. The percentage of learners gaining a level remains high for the final three categories (100–119, 120–139, 140 or more). What this descriptive data strongly implies is that hours of instruction do have a noticeable impact on gains in NRS EFLs, replicating the trend that Young (2007) reported.

However, because each NRS EFL corresponds to a band of BEST Plus scores, gain of a level might not always be indicative of a meaningful score gain. That is, with a cut score of 428 between the Low Beginning ESL and High Beginning ESL levels, a score of 425 on BEST Plus places one in the Low Beginning ESL level and a score of 430 places one in the High Beginning ESL level; a gain in level in such a case is demonstrated on the basis of a five-point improvement in the score. To gain insight on the amount of level gains that might be a result of score fluctuation due to measurement error in test scores, an additional descriptive analysis was conducted. First, a gain of 20 points on the BEST Plus was used as an indication of meaningful gain after accounting for measurement errors (20 points was chosen as an indicator of meaningful change on the BEST Plus scale score because it is equivalent to one standard error of measurement on the BEST Plus scale). The differences between pre-test and post-test scores were then examined to see how many learners attained score gains of greater than 20 points. The results are shown in Table 22.

Table 22. Number of Learners Attaining Meaningful Score Gain on BEST Plus by Instructional Hours

Instructional hours	Learners who attained meaningful gain
0 to 19 Hours n=41	21 (51%)
20 to 39 Hours n=160	124 (78%)
40 to 59 Hours n=320	270 (84%)
60 to 79 Hours n=1392	1196 (86%)
80 to 99 Hours n=460	410 (89%)
100 to 119 Hours n=225	198 (88%)
120 to 139 Hours n=126	116 (92%)
140 or More Hours n=236	204 (86%)
Total n=2960	**2539 (86%)**

A comparison of Table 22 and Table 21 shows that the percentage of learners in each instructional hour category is *higher* for the score gain (Table 22) than for the level gain (Table 21). This is encouraging, as it indicates that it is likely that few of the level gains seen in Table 21 are the result of minor score gains across a given cut score. In fact, it seems to indicate the opposite, that a number of learners are showing meaningful improvements that are not being captured by the level-gain category. To investigate these values further, the values for level gain and for meaningful score gain for each of the instructional hour categories were cross-tabulated.

Table 23 below shows the cross-tabulated breakdown (level gain by meaningful gain) for the group of learners who received 0–19 hours of instruction. Two values are of primary interest in this table. Looking across the top Meaningful Gain rows to the far right column, it can be seen that 51.2% of these learners showed at least a 20-point gain on the BEST Plus. Although that gain only translated into a level gain for 36.6% of the learners, it is arguably just as important to show score gain as level gain. The gain percentage of 51.2% is modest, but it should be kept in mind that this is the group receiving the fewest hours of instruction. It would be surprising for a larger majority of learners in this instructional hour category to show gains after such a small amount of instruction (unless their outside class activities involved high rates of English use and practice). The other value of interest in this table is the Level Gain and No Meaningful Gain cell in the first column under Level Gain. For the 0–19 hours group, it can be seen that this value is 0, meaning that 0% of this group demonstrated a level gain on the basis of a non-meaningful score gain.

Table 23. 0–19 Hours Group

		Level gain	No level gain	Total
Meaningful gain (20+ points)	Count	15	6	21
	% of Total	36.6%	14.6%	51.2%
No meaningful gain (<20 points)	Count	0	20	20
	% of Total	0.0%	48.8%	48.8%
Total	**Count**	**15**	**26**	**41**
	% of Total	**36.6%**	**63.4%**	**100.0%**

Table 24 shows the cross-tabulated values for the 20–39 instructional hours group. Here, looking at the far right column for Meaningful Gain, it can be seen that 77.5% of the learners fall into that category. So while 64.4% of the learners in this instructional hours group showed a level gain, more than three quarters were showing meaningful score improvement. Looking at the table's Level Gain and No Meaningful Gain cell, it can be seen that this value is very small; only one of these learners crossed a cut-score threshold with a non-meaningful score gain.

Table 24. 20–39 Hours Group

		Level gain	No level gain	Total
Meaningful gain	Count	102	22	124
	% of Total	63.8%	13.8%	77.5%
No meaningful gain	Count	1	35	36
	% of Total	0.6%	21.9%	22.5%
Total	**Count**	**103**	**57**	**160**
	% of Total	**64.4%**	**35.6%**	**100.0%**

The encouraging cross-tabulation totals seen in Table 24 were repeated in Table 25 and Table 26 for the groups of learners with 40–59 and 60–79 hours of instruction. Respectively, 84.4% and 85.9% of these learners showed meaningful score gain between their pre-test and post-test scores and only 2.5% and 2.4% of the learners showing a level gain were in the non-meaningful score gain category.

Table 25. 40–59 Hours Group

		Level gain	No level gain	Total
Meaningful gain	Count	238	32	270
	% of Total	74.4%	10.0%	84.4%
No meaningful gain	Count	8	42	50
	% of Total	2.5%	13.1%	15.6%
Total	**Count**	**246**	**74**	**320**
	% of Total	**76.9%**	**23.1%**	**100.0%**

Table 26. 60–79 Hours Group

		Level gain	No level gain	Total
Meaningful gain	Count	1041	155	1196
	% of Total	74.8%	11.1%	85.9%
No meaningful gain	Count	34	162	196
	% of Total	2.4%	11.6%	14.1%
Total	**Count**	**1075**	**317**	**1392**
	% of Total	**77.2%**	**22.8%**	**100.0%**

Tables 27-30, presenting the cross-tabulated values for the groups with 80–99, 100–119, 120–139, and greater than 140 hours of instruction, show even more encouraging results. Each of these tables shows close to or over 90% of the learners demonstrating meaningful score gain after instruction, whereas the number of learners in the level gain category who did not improve meaningfully with their scores remains very low (1.6–2.7%).

Table 27. 80–99 Hours Group

		Level gain	No level gain	Total
Meaningful gain	Count	357	53	410
	% of Total	77.6%	11.5%	89.1%
No meaningful gain	Count	10	40	50
	% of Total	2.2%	8.7%	10.9%
Total	**Count**	**367**	**93**	**460**
	% of Total	**79.8%**	**20.2%**	**100.0%**

Table 28. 100–119 Hours Group

		Level gain	No level gain	Total
Meaningful gain	Count	168	30	198
	% of Total	74.7%	13.3%	88.0%
No meaningful gain	Count	6	21	27
	% of Total	2.7%	9.3%	12.0%
Total	**Count**	**174**	**51**	**225**
	% of Total	**77.3%**	**22.7%**	**100.0%**

Table 29. 120–139 Hours Group

		Level gain	No level gain	Total
Meaningful gain	Count	103	13	116
	% of Total	81.7%	10.3%	92.1%
No meaningful gain	Count	2	8	10
	% of Total	1.6%	6.3%	7.9%
Total	**Count**	**105**	**21**	**126**
	% of Total	**83.3%**	**16.7%**	**100.0%**

Table 30. >140 Hours Group

		Level gain	No level gain	Total
Meaningful gain	Count	188	16	204
	% of Total	79.7%	6.8%	86.4%
No meaningful gain	Count	5	27	32
	% of Total	2.1%	11.4%	13.6%
Total	**Count**	**193**	**43**	**236**
	% of Total	**81.8%**	**18.2%**	**100.0%**

To summarize the results from the cross-tabulations, it is apparent that looking solely at level gain percentages can mask the improvements that some learners are making in terms of score gains. Additionally, high rates of learner score gains are evident in all the instructional-hour categories, but are particularly evident in the instructional-hour categories that go beyond the 19-hour mark. Beyond the 0–19 hour category, between 75% and 90% of the learners showed meaningful score gains. Much of these score gains are reflected in level gains as well (77% of the total sample showed level gain), but seeing the percentages showing meaningful score gain regardless of level gain is additional evidence that instruction seems to have a positive relationship with improvement as measured by the BEST Plus after 20 hours of instruction. In addition, after 20 hours of instruction, the percentage of learners who showed neither level nor score gain falls to 21.9% (from 48.8% with less than 20 instructional hours). After 40 hours, this rate falls to 13.1% and never rises above that with more hours of instruction. These results collectively suggest that BEST Plus is sensitive to improvement associated with instruction even after 20 hours.

A final set of analyses were carried out on the pre-test and post-test BEST Plus scores for the learners in each instructional-hour category. For each instructional-hour group, a paired sample t-test was conducted to investigate the statistical significance of the difference in the means of the pre-test and post-test scores. The results, shown in Table 31, correspond to the descriptive data discussed above. Except for the 0–19 group, the gains demonstrated by all instructional-hour groups were statistically significant. To the extent that more instructional hours lead to proficiency gains, BEST Plus seems to be sensitive to instruction in that the mean differences between pre- and post-test scores increase as instructional-hour categories increase.

Table 31. Results of Paired Sample T-Tests Comparing Pre-Test and Post-Score Score Means

Instructional Hour Group	Pre-test score mean	Post-test score mean	Difference in means	Degrees of freedom	Significance
0-19	404.80	435.68	30.88	40	0.391
20-39	391.20	453.99	62.79	159	<0.001
40-59	398.87	471.29	72.42	319	<0.001
60-79	403.22	476.10	72.88	1391	<0.001
80-99	407.28	484.60	77.32	459	<0.001
100-119	385.72	465.65	79.93	224	<0.001
120-139	391.60	474.87	83.28	125	<0.001
>=140	389.66	475.84	86.19	235	<0.001

Conclusion. Using the new cut scores and smaller increments of hours of instruction from those reported in Young (2007), the descriptive analyses reported above seem to provide clear answers to both research questions. First, the relationship between instructional hours and NRS EFL gain as demonstrated through BEST Plus scores is positive. About 77% of all learners who received instruction showed a level gain on their BEST Plus post-test. Secondly, the relationship between instructional hours and meaningful score gain on the BEST Plus is also positive. About 86% of the learners who received instruction showed a gain of at least 20 points per their BEST Plus scale score. High rates of improvement begin to be seen according to both level and score gain criteria once learners have crossed the 20-hours-of-instruction threshold. These data, therefore, support CAL's current recommendation of 60 instructional hours before retesting, and suggest the possibility that BEST Plus results could be meaningful after 40 or even 20 hours of attended instruction.

5. Operational Item Bank Health

5.1. Item exposure

A practical issue inherent to computer-adaptive testing is the potential for unbalanced or excessive use of items. That is, because items in a computer-adaptive test are typically drawn from the same item pool for multiple examinees over multiple administrations, some items may be administered at much higher rates than others. Particularly, the initial items presented to test-takers may be administered repeatedly if the entry level ability estimate is assumed to be the same for all test-takers. Additionally, if the sample of test-takers for a given test tends to cluster around a particular ability level, items with difficulty estimates best suited for that ability level may be used more often than others. The usage rate of particular items, also referred to as item exposure, can be a problem for a variety of reasons, including situations where test-takers sit for the exam on multiple occasions or situations where test-takers discuss an exam after the test. Items with high exposure rates could become well-known, which would then compromise the integrity of the item pool. Because BEST Plus is a performance-based assessment in which there are no "right" or "wrong" answers (i.e., it is the language examinees use to express themselves that is evaluated, not the content of the answers), the danger inherent with overexposure is somewhat lessened. However, advanced knowledge of what the questions are could lead to examinees rehearsing their responses prior to the assessment.

A number of different strategies have been developed to try to address the issue of item exposure. These strategies have been grouped into four categories: randomization, conditional selection, stratified strategies, and combined strategies (Georgiadou, Triantafillou, & Economides, 2007). BEST Plus utilizes a combined strategy approach, including randomization and conditional selection criteria, to control item exposure. These strategies and the results of descriptive analyses conducted to investigate their effectiveness are described below.

a. Computer-adaptive procedures for selecting items for BEST Plus administration

The BEST Plus item bank contains items with content that relates to twelve different domains (e.g., education, family/parenting, health). There are two to four folders of items for each domain, amounting to a total of 36 separate folders in the item bank. Each folder contains eight elements: an introductory statement (not scored) and seven scored items, one each from the seven different item types that are included in the test: entry question (EQ), photo description (PD), yes/no question (YN), choice question (CQ), personal expansion (PE), general expansion (GE), and elaboration (EL).

Every BEST Plus examinee starts the test with an estimated ability level, in Rasch logits, of 0. All examinees receive the same six warm-up items to start the test, which are scored and result in an initial examinee ability estimate. However, the first folder of items that examinees are tested from is selected completely at random from the 36 folders in the item pool. Examinees are administered four of the items from the randomly selected first folder. The first two items that they are administered are always the entry question and photo description items from within the selected folder. The next two items they receive are the two items (from the remaining five items in the folder) that have item selection difficulty estimates that best match the examinees' initial ability estimate. Note that the examinee's ability estimate is updated after each item is administered.

After four items from the first folder are administered, the examinee is then administered three items from a new folder. The selection of the second folder is both conditional and randomized. First, if, for example, the initial folder was from the health domain, the remaining health domain folders are no longer available; examinees never receive items from more than one folder per domain. From the domains that are available, folders are identified that contain an item (other than entry question or photo description) that is no greater or less than .75 logits of item selection difficulty distant from the examinee's current ability estimate. Then, from the available pool of appropriate folders, one of them is selected by random for the examinee. He or she is then administered three items from that folder. (Note: if no folders contain items other than entry question or photo description that are within .75 logits of the examinee's ability, which would mean that the examinee most likely has a very high or very low ability estimate, folders with the most difficult five items or easiest five items are identified and the random selection is made from them.)

Examinees can receive up to three folders with three items, using the same criteria as above for each of the three-item folder selections. However, the exam can be stopped at any point if either of these two stopping criteria are met: 1) the test-taker shows an ability estimate is below a scale score of 330 for six consecutive questions, or 2) the standard error of the test-taker's ability estimate is less than or equal to 0.20. (The third and final stopping rule is that tests terminate after examinees have received 25 items.)

If an exam continues after a test-taker has been administered three items from three folders (meaning that they been administered 19 items thus far: 6 warm up, 4 from the first folder, and 3 from three more folders), they are administered items in pairs from up to three additional folders. Again, folders from domains that have previously been administered are omitted from the selection process for the two-item folders. From the available domains, folders that contain an item (other than entry question or photo description) that is estimated to have an item selection difficulty of no more or less than 0.50 logits distance from the examinee's ability estimate are identified and one is randomly selected. As with the three-item folder process described above, if no folders are identified with items other than entry questions or photo descriptions that are within 0.50 logits of ability, the folders with the five easiest items (if the test-taker's ability estimate is low) or five most difficult items (if the test-taker's ability estimate is high) are identified and one is randomly selected.

Examinees can receive up to three folders with two items. The assessment will terminate if, after completing the two items in a folder, the standard error of the test-taker's ability estimate is less than or equal to 0.20. Otherwise, the assessment will continue until an examinee receives three folders with two items each. If this happens, it means they will have been administered 25 items and the test will be terminated at that point. Table 32 provides a summary of the BEST Plus item selection procedure.

Table 32. Summary of BEST Plus administration

Phase	# of Folders*	# of Questions*	Path	Scored	Folder Selection
Sign-in				No	N/A
Welcome				No	N/A
Warm-Up	1	6	Fixed	Yes	Same for all examinees
4-item folder	1	4 per folder	Adaptive	Yes	Completely at random
3-item folders	3	3 per folder	Adaptive	Yes	Selected at random from • folders in domains not yet administered AND • folders containing an item other than EQ or PD whose selection difficulty is within 0.75 logits of the examinee's ability estimate • OR if no folders with non-EQ/PD item within 0.75 logits are identified, folder is selected at random from folders containing the five easiest (for low ability examinees) or five most difficult (for high ability examinees) items
2-item folders	3	2 per folder	Adaptive	Yes	Selected at random from • folders in domains not yet administered AND • folders containing an item other than EQ or PD whose selection difficulty is within 0.50 logits of examinee's ability estimate • OR if no folders with non-EQ/PD item within 0.50 logits are identified, folder is selected at random from folders containing the five easiest (for low ability examinees) or the five most difficult (for high ability examinees) items
Wind-Down (WD)				No	

*Test can stop prior to administering maximum of 25 items if a) the standard error estimate of the examinee's ability after an item is less than 0.20 or b) if after the warm-up or after six items in a row the examinee's ability estimate is less than a scale score of 330.

b. Item exposure on BEST Plus

Through the randomization and conditional selection strategies described above, the BEST Plus design allows for a degree of control over item exposure. The 2011-2013 operational data set was used to investigate how effective these strategies are.

Table 33 provides a summary of the distribution of NRS EFLs within the sample. This analysis used the cut scores established during the 2012 standard-setting study. A large proportion of the sample was at the lower levels, with slightly less than half of test-takers at levels 1 and 2 (47.7%). However, there is a fairly even distribution across levels 3–7, so it can be said that the sample covers the entire range of ability. In addition, this sample resembles the broader population of BEST Plus test-takers; in general, more test-takers are at lower ability levels than higher ability levels.

Table 33. Distribution of NRS EFLs in operational sample

NRS EFLs*	Frequency	Percent	Cumulative Percent
Level 1	245	18.1	18.1
Level 2	401	29.6	47.7
Level 3	155	11.5	59.2
Level 4	189	14.0	73.2
Level 5	146	10.8	84.0
Level 6	89	6.6	90.5
Level 7 (exit from NRS EFLs)	128	9.5	100.0
Total	**1353**	**100.0**	

*For ease of reporting, this section uses the following key to represent the NRS EFLs: Beginning ESL Literacy = 1, Low Beginning ESL = 2, High Beginning ESL = 3, Low Intermediate ESL = 4, High Intermediate ESL = 5, Advanced ESL =6, Exit from NRS EFLs = 7.

The rates of item exposure for this sample of BEST Plus test administrations were looked at in three ways: conditional on NRS EFL, conditional on item type, and per item.

Table 34 includes descriptive data on rates of exposure for items by NRS EFL (columns) and item type (rows). The first "Sum" row shows the number of entry question (EQ) items administered per NRS EFL (780 for NRS EFL 1, 1,229 for NRS EFL 2, etc.) and the final column shows the total number of EQ items administered for this sample (k==5,476). The "Sum" rows for PD (photo description), YN (yes/no question) and CQ (choice question) show the number of those items administered per NRS EFL. The more difficult items, personal expansion (PE), general expansion (GE), and elaboration (EL), were treated differently from the four easier item types, as these more difficult items are designed to elicit more extended responses. Because these items were administered less frequently (both by design—as described above, the item selection algorithm is set up to deliver a certain number of PD and EQ items as they are always the first item administered in a folder, PD particularly for lower ability examinees—and by virtue of the ability of the test-takers in this sample being on the low end of the scale), the three item types were grouped into three categories of equal size ("low difficulty," "medium difficulty," and "high difficulty") to better capture their exposure.

Table 34. Exposure rates by NRS EFL and item type

Item type		**NRS EFL 1**	**NRS EFL 2**	**NRS EFL 3**	**NRS EFL 4**	**NRS EFL 5**	**NRS EFL 6**	**Level 7 (exit)**	**Total**
EQ*	Sum	780	1229	554	753	733	568	859	5476
	row %	14%	22%	10%	14%	13%	10%	16%	29%
	col%	25%	26%	27%	28%	30%	34%	35%	
PD*	Sum	1031	1379	418	366	258	130	165	3747
	row %	28%	37%	11%	10%	7%	3%	4%	20%
	col%	33%	30%	20%	13%	11%	8%	7%	
YN*	Sum	766	768	309	306	141	31	9	2330
	row %	33%	33%	13%	13%	6%	1%	0%	12%
	col%	24%	16%	15%	11%	6%	2%	0%	
CQ*	Sum	508	1005	389	406	166	34	8	2516
	row %	20%	40%	15%	16%	7%	1%	0%	13%
	col%	16%	22%	19%	15%	7%	2%	0%	
L**	Sum	76	270	296	660	762	466	319	2849
	row %	3%	9%	10%	23%	27%	16%	11%	15%
	col%	2%	6%	14%	24%	32%	28%	13%	
M**	Sum	0	18	80	195	278	313	510	1394
	row %	0%	1%	6%	14%	20%	22%	37%	7%
	col%	0%	0%	4%	7%	12%	19%	21%	
H**	Sum	0	0	19	42	74	117	562	814
	row %	0%	0%	2%	5%	9%	14%	69%	4%
	col%	0%	0%	1%	2%	3%	7%	23%	
	Total	**3161**	**4669**	**2065**	**2728**	**2412**	**1659**	**2432**	**19,126**

*EQ=Entry Question, PD=Photo Description, YN=Yes/No, CQ=Choice Question

**L=lowest 1/3 difficulty grouping, M=medium 1/3 difficulty grouping, H=highest 1/3 difficult grouping of aggregate of Personal Expansion, General Expansion, and Elaboration items

The EQ item type is the most evenly administered across all NRS EFLs and its exposure rate increases slightly as the NRS EFL increases. This is by design as the entry question item type is administered as the first item in a folder, except that lower performing test-takers, who may not understand an EQ, may be presented a PD item as the first item in a folder. As expected, its use ranges from 25% of all 3,161 items (see bottom of NRS 1 column) being administered to NRS EFL 1 test-takers to 35% of all 2,432 items being administered to NRS EFL 7 test-takers.

For NRS EFL 1 test-takers, the greatest percentage of their items (33%) were photo description items and the next greatest percentage (24%) were Yes/No items. Photo description was also the most frequent item type for NRS EFL 2 (30%) and NRS EFL 3 (20%) and for both of those levels, the choice questions were the second most frequent, at 22% and 19% respectively. These higher rates of use of the easier items (photo description, yes/no questions, and choice questions) are in line with the adaptive nature of the test. We expect to see lower ability test-takers being administered the easier items at higher rates and the more difficult items at lower rates (e.g., NRS EFLs 1, 2, and 3 were rarely administered the expansion and elaboration type questions, although, as expected, level 3 received more than level 2, which received more than level 1). At NRS EFL 4, the middle level of ability in the sample, the pattern begins to change. The item type administered at the greatest rate to NRS EFL 4 test-takers is the low-difficulty category of the more difficult item types (24%), but the second highest item type is an easier one, choice questions (15%), which is reflective of the notion of NRS EFL 4 straddling the beginner-level NRS categories and the higher-level categories where test-takers can respond at greater length to questions. Test-takers with the ability to respond at greater length to items are those in NRS EFLs 5 and above. Levels 5 and 6 are administered the low-difficulty expansion and elaboration items at the highest rate (32% and 28%) and the medium-difficulty ones at the second highest rate, 12% and 19% respectively. Finally, for the highest level, level 7, the test-takers receive the most challenging of the expansion and elaboration items at the highest rate (23%) and the medium-difficulty difficult items at the second highest rate (21%). As expected, examinees at the highest levels (NRS 5, 6, and 7) are administered the easier items (photo description, yes/no questions, and choice questions) at very low rates, particularly for yes/no and choice questions, which range from 0% to 7% of their totals (photo description being slightly higher as a result of photo description items being required in the first folder for all test takers).

The exposure rates by item type reveal a similarly expected pattern. Photo description, yes/no, and choice items were administered at the highest rate to NRS EFL 2 test takers and at the second highest rate to NRS EFL 1 test takers. Considering the high number of NRS EFL 2 test-takers in the sample, this pattern makes sense. The highest-difficulty and medium-difficulty expansion and elaboration items were administered at their greatest rates to the NRS 7 test takers and their second highest rate to the NRS EFL 6 test takers. The low difficulty expansion and elaboration items were administered at their highest rate to NRS 5 test-takers and second highest rate to NRS 4 test takers.

Table 35 shows the descriptive data on item selection difficulty for the seven item types. Collectively, the easier item types (entry questions, photo description, choice questions, and yes/no questions) are performing as expected, all with negative average item selection difficulty measures. The more difficult item types (personal expansion, general expansion, and elaboration) are also performing as expected, all with positive item selection difficulty measures. Returning to item exposure rates by item types and NRS EFLs, when rows are ranked from easiest items to most difficult and columns ascend in ability level, this somewhat diagonal pattern of exposure rates, conditional on NRS EFL and item type, is what is expected for a computer adaptive test that is working properly. Additionally, because of the randomization of the design, and forced administration of some item types, we only see little or no usage in the bottom left corner (difficult item types were not administered to low level ability test-takers) and on the far right edge (very few easy item types were administered to high level ability test-takers).

Table 35. Average item selection difficulty measures by item type

Item type	Mean	Number of items	Standard Deviation
Entry question	-0.36	36	0.69
Photo description	-1.35	36	0.20
Choice question	-1.03	36	0.43
Yes/no	-1.18	36	0.56
Personal expansion	0.63	36	0.52
General expansion	0.94	36	0.54
Elaboration	1.01	36	0.46

The data in Table 34 provide a picture of how the different item types are being exposed across levels, but an examination of the exposure of individual items (a total of 258 items) is also required. Complete item-level exposure information is in Appendix A. To summarize the item-level data, leaving aside the six warm-up items, which are administered to all test-takers and therefore have 100% exposure rates, the highest rate of exposure for any individual item is 18% (meaning 18% of the total sample of 1,353 test takers were administered the item). This is not an inordinately high rate of exposure, and we also see that only three items were exposed at that rate (or only about 1% of the total item bank) and only 11 items (roughly 4% of the item bank) were exposed to 15–18% of the sample. Another 27 items were exposed to 10–14% of the population. The vast majority of the item bank, or 82% (213 items), were administered to 1–9% of the sample. These item-level exposure rates indicate that the randomization and conditional strategies being employed by the BEST Plus selection algorithm are effective in controlling item exposure. Examinees are being exposed to items that are appropriate to their ability level. This finding ensures that, as examinees are retested, to the degree that they are progressing in their acquisition of oral language proficiency in English, they will be exposed to different types of items in the pool that are appropriate to their level of oral language proficiency.

The item-level data also provide evidence that the adaptive nature of the test is working. Among the high exposure items, we see primarily easy item types. This is expected as all examinees are tested with a certain number of easier items and lower-level examinees are tested with more of them. Also, as Table 33 showed, there were more lower-level ability test-takers in the sample than higher-level ability; a little less than 50% of the examinees were in the two lowest NRS EFLs.

Further, the 15 individual items with the greatest rates of use (or highest exposure) were all either entry question or photo description items. For example, entry question item "hous.2.eq", which has an exposure rate of 18%, has an estimated item selection difficulty on the easy side, of -0.96 logits. It was administered a total of 249 times. These administrations were spread out across the NRS EFLs but concentrated at NRS EFLs 1 and 2, with roughly 39% of its use occurring at those levels. A similar pattern is seen with the photo description item "old.fami.3.PD," which was administered to 16% of the total sample and has an item selection difficulty of -1.28 logits. This was an easier item than "hous.2.EQ" and was administered at a greater rate to NRS EFL 1 and 2 test takers.

At the other end of the scale, a similar picture is evident; the more difficult items (personal expansion, general expansion, and elaboration) are administered less often and have higher difficulty estimates. Of the 31 items that were administered to roughly 1% of the sample, 26 were personal expansion, general expansion, or elaboration items. For example, personal expansion item "gett.3.PE" was administered to only 1% of the total sample. This corroborates with its higher item selection difficulty estimate of 1.23 logits. When it was used (a total of 8 times) it was solely with the NRS EFL 6 and 7 test takers. The elaboration item "civi.2.EL" has a similar pattern. With a difficulty estimate of 1.06 logits, it was administered a total of 19 times (1.4% of the sample) and 13 of those uses were with the highest level and its other 6 uses evenly split between NRS EFL 5 and NRS EFL 6.

There are some anomalies, however, which are also enlightening to examine. For example, items “fami.1.YN” and “fami.1.CQ” were easy item types that were administered at very low rates; however, a close examination of their item selection difficult revealed that they were actually more difficult than other of the easier items. Their selection difficulty, -0.58 and -0.50 logits, are higher than the average selection difficulty for yes/no and choice question item types at -1.18 and -1.03 logits, respectively. On the other end of the spectrum, items “weat.2.GE” and “recr.3.PE” were difficult item types administered at high rates. Closer examination reveals that these two items were in fact easy expansion and elaboration items. Their selection difficulty, -0.11 and -0.64 logits, are lower than the average selection difficulty for general expansion and personal expansion item types at 0.94 and 0.63 logits, respectively. Regardless of the anomalies in terms of item types, we see expected exposure patterns across the NRS EFLs, such that items with lower selection difficulty were more likely to be administered to lower NRS EFL test takers and that items with higher selection difficulty were more likely to be administered to higher NRS EFL test takers. The overall picture supports the claim that the selection strategies for the BEST Plus are working effectively and as designed. A study of the anomalies gives insights into tightening the specifications for the development of future BEST Plus items.

c. Summary

A descriptive analysis of item exposure in the BEST Plus reveals that the item exposure control strategies are effective in ensuring a desirable spread of item type use by item type and examinee ability level and preventing individual items from being used at inordinate rates. The combination of an adequately sized item bank and randomization and conditional selection rules in the item selection algorithm have been shown to be useful strategies for preventing undesirable item exposure. Additionally, given the effective exposure control, item exposure does not appear to have an effect on item drift of individual items (see next section for detail).

5.2. Item drift

An important assumption underlying computer-adaptive tests such as the BEST Plus is that item parameter estimates calibrated with one sample remain stable, or invariant, across different samples from different test administrations. Item parameter change across test administrations is known as item parameter drift (Goldstein, 1983) and the degree to which it exists should be investigated for items that are used in multiple test administrations. If there has been a great deal of change in an item's parameters, the contribution of performances on that item to examinees' measurement may lead to distortion. If drift happens across many items, interpretations made on the basis of examinees' performances on those items may no longer be valid.

In order to investigate item parameter drift for the BEST Plus, item parameter estimates from two different recent data sets were examined in relation to the items' original 2003 parameter estimates. The two data sets used to investigate item drift were:

1. Responses from operational administrations (n=1,353) of the BEST Plus carried out by adult ESL programs in different parts of the United States during program years 2011–2013. Responses from 258 items were included in the dataset: the 168 items that will remain in the item bank going forward and the 90 items that will be replaced by new items.
2. Responses from the 2012 field test (n=2,961) of the BEST Plus, in which the new items were embedded. Responses from the 168 continuing items were included. Responses from the 90 new items being field tested were not included; by virtue of being new, they did not have existing estimates from which they could vary.

Table 36 provides a breakdown of the items that were included in the current analyses for item drift.

Table 36. Items included in item parameter drift analyses

Operational items (2011–2013)	168 continuing items	90 (to be retired, included in the analysis)
Field test items (2012)	168 continuing items	90 (new, not included in the analysis)

Although the 90 items to be retired were included in the estimates of item difficulty for the operational items data set, because they are being removed from the item bank, they were not themselves examined for item drift. The focus of the item drift analyses was on the 168 continuing items and the point of reference for potential drift was their original item estimates from when the BEST Plus was initially constructed in 2003. Combining the two datasets resulted in a sample size of 4,314 test takers. Table 37 shows the number of items (column 2) by the number of responses the items received (column 1). Because the BEST Plus is computer adaptive, not all test-takers respond to the same number of items in an administration, and as a result the number of overall responses for items varied. As shown in Table 37, in the combined data set (n=4,314), 30.4% of the 168 continuing items received 0–99 responses, 21.4% received 100–199 responses, 14.3% received 200–299 responses, and 33.9% received more than 300 responses.

Table 37. Number of items by number of responses per continuing item

Number of responses per item	Number of items	Percent	Cumulative percent
0–99	51	30.4	30.4
100–199	36	21.4	51.8
200–299	24	14.3	66.1
300 or more	57	33.9	100.0
Total	**168**	**100.0**	

To screen for item drift among the 168 continuing items, two statistics were used. First, the displacement statistic that is generated by Rasch analyses was examined for each continuing item. The displacement statistic "approximates the displacement of the estimate away from the statistically better value which would result from the best fit of your data to the model" (Linacre, 2012). In addition, the Robust Z-statistic (Huynh & Meyer, 2010; Kim, Baron, & Choi, 2010) was calculated. The rationale for using the two statistics was to see if there was a convergence of results for the items demonstrating item drift.

The Rasch displacement value is automatically calculated for any Rasch-Grouped Rating Scale analysis that includes anchored items. To obtain the Rasch displacement index using the current dataset (N=4,314), an anchored run was carried out (using FACETS version 3.71.4) for all 258 items using the original 2003 item difficulty parameters and scale step parameters. Items with displacement > |0.5| were flagged for item drift. Keeping in mind that only the 168 continuing items are the focus of this study, only 9 of those 168 continuing items (about 5.4%) were flagged for displacement (indicated in the "Yes" column under the "Displacement header" in Table 38). Table 38 shows that four of the flagged items were in the 0–99 response category, two of the flagged items were in the 100–199 response category, one flagged item was in the 200–299 response category, and two of the flagged items were in the 300 or more response category. Details on the items' displacement values and item type will be provided below.

Table 38. Displacement by item-response categories (168 continuing items)

		Displacement?		
		No	**Yes**	**Total**
0–99 responses per item	Item Count	47	4	51
	% within category	92.2%	7.8%	100.0%
	% of total	28.0%	2.4%	30.4%
100–199 responses per item	Item Count	34	2	36
	% within category	94.4%	5.6%	100.0%
	% of total	20.2%	1.2%	21.4%
200–299 responses per item	Item Count	23	1	24
	% within category	95.8%	4.2%	100.0%
	% of total	13.7%	0.6%	14.3%
300 or more responses per item	Item Count	55	2	57
	% within category	96.5%	3.5%	100.0%
	% of total	32.7%	1.2%	33.9%
Total	**Item Count**	**159**	**9**	**168**
	% of total	**94.6%**	**5.4%**	**100.0%**

As mentioned earlier, a second statistic, the Robust Z-statistic, was estimated to screen for items showing invariant performance across test administrations. Researchers (e.g., Huynh & Meyer, 2010; Kim et al., 2010) have advocated using a statistic such as the Robust Z-statistic, which uses median item difficulty rather than mean item difficulty, in conjunction with (or in place of) a mean-based statistics such as the displacement index because mean-based statistics can be susceptible to extreme values. The equation for the Robust Z-statistic (Kim et al., 2010) is shown in Equation 1:

Equation 1. Robust Z-statistic calculation

$$Z_i = \frac{\left(b_{iF} - \hat{b}_{IE}\right) - MD_d}{0.74 \times INQ_d}$$

In this equation b_{iF} is the fixed item parameter for a given item, which in this analysis is the original item bank difficulty estimate. The new freely estimated item parameter for the item, using the responses in the current (2011–2013) dataset, is represented by $\hat{b}_{IE}$. MD_d represents the median of the differences between the fixed difficulty parameters and freely estimated difficulty parameters across all the items. INQ_d represents the interquartile range of the differences between the fixed parameters and the freely estimated parameters. Using a 95% confidence interval, any item whose Robust Z value was equal to or larger than 1.96 was flagged for drift. As shown in Table 39, only 6 (3.6%) of the 168 continuing items were flagged for drift using the Robust Z-statistic (indicated in the "Yes" column under the "Robust Z" header). As with the displacement index results, the greatest number of flagged items (four) was in the 0–99 response category. Only one item was flagged in both the 100–199 response category and 200–299 response category, and none of the items in the 300 or more response category were flagged for drift.

Table 39. Robust Z-statistics by item-response categories (168 continuing items)

		Robust Z		
		No	**Yes**	**Total**
0–99 responses per item	Item Count	47	4	51
	% within category	92.2%	7.8%	100.0%
	% of total	28.0%	2.4%	30.4%
100–199 responses per item	Item Count	35	1	36
	% within category	97.2%	2.8%	100.0%
	% of total	20.8%	0.6%	21.4%
200–299 responses per item	Item Count	23	1	24
	% within category	95.8%	4.2%	100.0%
	% of total	13.7%	0.6%	14.3%
300 or more responses per item	Item Count	57	0	57
	% within category	100.0%	0.0%	100.0%
	% of total	33.9%	0.0%	33.9%
Total	**Item Count**	**162**	**6**	**168**
	% of total	**96.4%**	**3.6%**	**100.0%**

The items that were flagged by each analysis were then examined to see which types of items they were and how much agreement there was between the two analyses. Table 40 shows the item count, displacement value, item name, item type, and domain for the nine items that were flagged for drift using the displacement index. A positive displacement value suggests that an item may have become more difficult based on the current calibration, whereas a negative displacement value is associated with an item becoming potentially easier over test administrations (Linacre, 2012). Table 41 shows the item count, Robust Z value, item name, item type, and domain for the six items that were flagged for drift using the Robust Z-statistic. Given the aforementioned equation of Robust Z, a positive Z value is associated with an item becoming potentially easier over test administrations, whereas a negative Z value suggests that an item may have become more difficult based on the current calibration.

Table 40. Items flagged for drift by displacement index

Count	Displacement	Name	Item Type	Domain
0-99	-0.50	educ.2.GE	General expansion	Education
0-99	1.14	recr.2.EL*	Elaboration	Recreation
0-99	-0.78	hous.3.GE	General expansion	Housing
0-99	-0.54	weat.1.EL	Elaboration	Weather
100-199	0.52	gett.1.PD	Photo description	Getting a job
100-199	-1.06	on-t.3.YN*	Yes/no	On the job
200-299	-0.76	tran.1.YN*	Yes/no	Transportation
300 or more	-0.81	warmup.2.WA	Warm-up	Warm-up
300 or more	-0.63	warmup.6.WA	Warm-up	Warm-up

*flagged in both analyses

Table 41. Items flagged for drift by Robust Z-statistic

Count	Robust Z	Name	Item Type	Domain
0-99	-3.10	recr.2.EL*	Elaboration	Recreation
0-99	-2.20	comm.1.GE	General expansion	Community services
0-99	-2.59	civi.1.GE	General expansion	Civics
0-99	-2.20	civi.1.EL	Elaboration	Civics
100-199	3.13	on-t.3.YN*	Yes/no	On the job
200-299	1.97	tran.1.YN*	Yes/no	Transportation

*flagged in both analyses

One very positive result of these analyses is to discover that very few items, even after several years of exposure, were flagged for drift in each analysis. Across the two different methodologies, there was agreement only on 3 of the 168 items (less than 2%). Among the six items flagged by Robust Z statistics, there seems to be a relationship between item types and the direction of item drift, such that several easy items (e.g., yes/no items) appear to have become even easier while several difficult items (e.g., elaboration and general expansion items) appear to have become more difficult. However, the picture was less clear among the nine items flagged by the displacement index. While most of the easy item types (e.g., yes/no, warm-up, and photo description) appear to have become easier, there was one exception (get.1.PD). In this analysis, however, most of the difficult item types (e.g., elaboration and general expansion) also appeared to become easier, again with one exception (recr.2.EL). The three items flagged by both analyses naturally show the same trend of those flagged by the Robust Z-statistic alone.

Finally, there were no discernible patterns among this group of flagged items by item type or folder. Items flagged for drift come from those that are more challenging (e.g., elaboration and general expansion) and those that are easier (e.g., warm-up questions, yes/no, photo description). Additionally, the 12 total flagged items were from 10 different folders. The cause of the drift, then, appears to be idiosyncratic. (A complete list of the 168 continuing items and their displacement values and Robust Z-statistics can be found in Appendix B.)

To sum up, two different statistics were used to identify items displaying item parameter drift (i.e., changes in item difficulty over time). The analyses identified only 12 items with potential drift issues (9 with the displacement index, 6 with the Robust Z-statistic, and 3 in common across the two analyses). This indicates that for the most part, difficulty estimates for the continuing items in the item bank have been very stable over the many years of the operational testing program. Maintenance of the item parameters in the item pool for a computer-adaptive test is an important aspect of maintaining the validity of the BEST Plus score results. More importantly, the current analysis serves to inform CAL's ongoing effort in maintaining the health of the BEST Plus item bank. The 12 items flagged in this analysis will be one of CAL's priorities in the next phase of supporting item-bank integrity by either replacing these items with new items or updating their item difficulty parameters.

CAL also undertook a study investigating the relationship between item exposure and item parameter drift. The focus of the item drift analysis was on the 168 continuing items that will remain in the updated item bank moving forward. Thus, the current study aimed to investigate whether there was a relationship between item exposure rates and item drift among the 168 continuing items. It is important to investigate this because items that are more exposed may be compromised by their familiarity and thus become easier with the passage of time.

Using exposure counts per item (from Appendix A) as item exposure in the 2011-2013 operational data set (N=1,353), CAL staff categorized item exposure into four groups with ascending item exposure: 123 of the 168 continuing items had 0–99 exposure counts, 32 had 100–199 counts, 7 had 200–299 counts, and 6 had more than 300 counts. Table 42 shows the number of items (column 2) by exposure counts the items had (column 1).

Table 42. Number of items by exposure counts per continuing item

Exposure counts per item	Number of items	Percent	Cumulative percent	
0–99	123	73.2	73.2	
100–199	32	19.0	92.3	
200–299	7	4.2	96.4	
300 or more	6	3.6	100.0	
Total	168	100.0		

For the 168 continuing items, those that were flagged for item drift by the displacement statistic and/or by the Robust Z-statistic were coded as "1" and those not flagged by the two statistics were coded as "0," creating two binary variables for each item in relation to item drift using the two statistics.

Because the exposure-count categories are ordinal, an ordinal test of independence was conducted between item exposure and item drift. When there is a true association between the ordinal variable and the binary variable, an ordinal test using M^2 statistic (df=1) is more sensitive in detecting the association (Agresti, 1996). For the binary variable created using the displacement statistic, the M^2 statistic was not significant ($M^2 = 1.447, df = 1, p = 0.229$), suggesting that there is no strong evidence of an association between item exposure and item drift. Similar, for the binary variable created using the Robust Z-statistic, the M^2 statistic was not significant ($M^2 = 1.683, df = 1, p = 0.195$). Thus, there is no evidence that more exposure is related to BEST Plus items becoming easier.

5.3. Item overlap

The selection of items for administration is particularly important in order to control for item overlap between a pre- and post- test administration. On the computer-adaptive version of BEST Plus, neither the test administrator nor the examinee knows in advance which items will be administered. While it is possible for an examinee to receive the exact same questions in the exact same order on two administrations of the computer-adaptive version, this is highly unlikely. When the print version of BEST Plus is used, the administrator is responsible for ensuring that each examinee receives different forms for the pre- and post-test (per the Test Administrator Guide), and therefore there should be no overlap between test administrations for the print-based version. However, the amount of item overlap between two administrations of the computer-adaptive version of BEST Plus can be empirically investigated.

In the fall of 2014, CAL conducted a study in which 33 examinees took the computer-adaptive version of BEST Plus 2.0 in one room and then were immediately retested in another room. Overlap was examined from two perspectives: folder overlap and item overlap. Twenty-two of the 33 examinees (67%) experienced some degree of item overlap. Table 43 shows the rates of overlap of folders between the two test administrations in this study. (Note: The data in the tables in this section do not include the warm-up items because these are intentionally the same for all examinees.)

Table 43. Rates of folder overlap between two test administrations

Number of Folders That Overlapped	Number of Examinees	% of Total Examinees (N=33)
0	11	33%
1	10	30%
2	10	30%
3	2	6%
Total with Overlap	**22**	**67%**

At the folder level, 11 examinees (33%) received no common folders across the two administrations, ten examinees (30%) received one folder in common, ten examinees (30%) received two folders in common, and two examinees (6%) received three folders in common across the two administrations. These results demonstrate that it is most likely that an examinee would have no or at most one folder in common across two test administrations (63% of examinees in this study) given back-to-back.

Table 44 shows the rates of overlap of items and folders across which the overlapping items are dispersed between test administrations in this study.

Table 44. Rates of item overlap between two test administrations

Number of Items That Overlapped	Number of Examinees	% of Total Examinees (N=33)
0	11	33%
1	1	3%
2	6	18%
3	4	12%
4	8	24%
5	2	6%
6	0	0%
7	1	3%
Total with Overlap	**22**	**67%**

At the item level, 11 examinees (33%) received no items in common, one examinee (3%) received one item in common, six examinees (18%) received two items in common, four examinees (12%) received three items in common, eight examinees (24%) received four items in common, two examinees (6%) received five items in common, and one examinee (3%) received seven items in common across the two administrations. These results suggest that it is most likely than an examinee would have no or at most three items in common across two test administrations (66% of examinees in this study) given back-to-back.

An additional analysis was conducted using only data from examinees who experienced some degree of item overlap (22 out of 33 examinees) from the current study. Table 45 presents the descriptive statistics for the 22 examinees across the two computer-adaptive administrations, in which there was some degree of item overlap. It demonstrates the similarity of mean performance in terms of scale scores across the examinees between the first administration of BEST Plus (Test 1) and the second administration of the test (Test 2).

Table 45. Descriptive statistics for computer-adaptive test administrations with item overlap

	N	Mean	Std. Deviation	Min	Max
Computer-adaptive Test 1	22	497.64	69.55	380	659
Computer-adaptive Test 2	22	492.18	61.70	376	619

As could be expected, the correlation between the two performances was high (.85). However, as seen in Table 45, average performances between the first and second administration did not increase. Indeed, they showed a slight decrease. Table 46 presents the scale score difference between the two computer-adaptive test administrations with item overlap. To examine whether there was a statistically significant difference between scale scores on the two computer-adaptive

test administrations with item overlap, CAL staff conducted a paired-sample t-test. The results, presented in Table 46, show that the difference between performances on the two test administrations was not statistically significant, suggesting that the administration of a few overlapping items ultimately does not have an impact on observed scores.

Table 46. Paired difference of scale scores for computer-adaptive test administrations with item overlap

	Paired-difference mean	***t***	***df***	***P***
Computer-adaptive Test 1 vs. Computer-adaptive Test 2	5.46	.69	21	.50

It is important to note that the situation in the above test-retest study does not reflect operational test administration at the program level. Programs do not, and should not, administer BEST Plus as a post-test immediately after the pre-test. Nevertheless, the study was for research purposes and the results serve to inform testing programs in terms of the situation in which the rate of item overlap would mostly likely have been the largest across computer-adaptive test administrations (i.e., no growth in student proficiency between administrations).

To examine the degree of item overlap in operational conditions, CAL staff also looked at the degree of item overlap from program data using the 2011-2013 operational data set. Among this operational sample, 500 examinees had complete pre- and post-test records, and therefore were included in the analysis for item overlap.

Table 47 and Table 48 show the rates of overlap by the number of folders and by the number of items, respectively.

Table 47. Rates of folder overlap for pre- and post-administrations in operational programs

Number of Folders That Overlapped	Number of Examinees	% of Total Examinees (N=500)
0	248	50%
1	170	34%
2	63	13%
3	13	3%
4	6	1%
Total with Overlap	**252***	**50%**

Table 48. Rates of item overlap for pre- and post-administrations in operational programs

Number of Items that Overlapped	Number of Examinees	% of Total Examinees (N=500)
1	41	8%
2	69	14%
3	75	15%
4	28	6%
5	20	4%
6	6	1%
7	5	1%
8	2	<1%
9	0	<1%
10	2	<1%
11	1	<1%
Total with Overlap	**249***	**50%**

*Note: the difference between rates of overlap at the folder and item levels suggests that three examinees received folders that overlapped between test administrations, but the items drawn from the folders were entirely different.

As expected, rates of overlap were less in the operational program where there was a time lapse (and potentially growth in student proficiency) than when the two administrations occurred back-to-back. In the operational data, at the folder level, 248 examinees (50%) received no folders in common, compared to only 33% in the back-to-back study. In the operational programs, 170 examinees (34%) received one folder in common, 63 examinees (13%) received two folders in common, 13 examinees (3%) received three folders in common, and 6 examinees (1%) received four folders in common.

At the item level, 251 examinees (50%) received no items in common, whereas, again, only 33% received no items in common when two administrations of BEST Plus occurred back-to-back. In the operational data, 41 examinees (8%) received one item in common, 69 examinees (14%) received two items in common, 75 examinees (15%) received three items in common, 28 examinees (6%) received four items in common, 20 examinees (4%) received five items in common, and about 3% of the examinees received six or more items in common. Overall, these results demonstrate that in operational testing, it is most likely that an examinee would have no or at most one folder in common (84% of examinees in the operational data) and would have no or at most three items in common (87% of examinees) across computer-adaptive test administrations of the BEST Plus.

Based on the research presented here, item exposure does not appear to be a significant influence on BEST Plus performances. Although there was a higher rate of item exposure, as expected, in the two BEST Plus administrations given back-to-back in the Fall 2014 study than in the operational program, that level of item exposure did not appear to influence average test performance. Thus, it is unlikely to be a factor given the lower rates of item exposure that occur in operational testing when the computer-adaptive BEST Plus is used as a pre-test and a post-test.

6. Interpreting Test Results

6.1. Standard-setting study

In 2012, CAL conducted a standard-setting study in order to establish new cut scores linking examinee results on BEST Plus to the NRS EFLs for English as a second language. Originally, CAL conducted a standard-setting study in 2002 that established cut scores linked to an earlier proficiency scale for adult English language education, the Student Performance Levels (SPLs). At the time, the SPLs had already been linked to the NRS EFLs, which ensured a connection between BEST Plus cut scores and the NRS EFLs. In 2006, the NRS revised the educational functioning levels for adult ESL. The highest level, High Advanced ESL, was merged with the Low Advanced ESL level and the Beginning ESL level was divided into two different levels: Low Beginning ESL and High Beginning ESL. Therefore, in 2012, CAL conducted a new standard-setting study in order to directly connect examinee results on BEST Plus to the revised NRS EFLs. The impetus for this study was clear alignment to the updated NRS EFLs with the link being made directly, rather than indirectly through the SPLs. Thus, the 2012 standard-setting study was a critical step in ensuring that BEST Plus remains a useful tool for measuring the English speaking proficiency of adults enrolled in ESL programs, serving the needs of those programs and those individuals.

Following technically sophisticated standard-setting procedures, outlined below, performances on BEST Plus were related to the ESL educational functioning level descriptors of the National Reporting System (NRS) for Adult Education. After the 2012 study was completed, an External Advisory Board (EAB) met in September 2013 to review the proposed changes to the cut scores and to vote on their adoption. The EAB accepted the new cut scores and voted for a July 1, 2015 adoption date.

The following sections describe in detail the procedures, judges, materials, and analysis of the 2012 standard-setting study. Then, the process and results of a 2013 meeting of the External Advisory Board (EAB) to vote on acceptance of the new scale scores are reported. Finally, a rationale is presented for the acceptance of the cut scores established during the 2012 standard-setting study.

6.1.i. Materials

In the 2012 standard-setting study, judges provided ratings for a selection of 21 student portfolios (i.e., full test administrations that were videotaped). Because the final NRS EFL cuts were to be expressed in terms of scale scores, BEST Plus scale scores were used in selecting representative samples. Because the NRS EFL descriptors are applied to adult proficiency in general and not separately to speaking and listening, the total BEST Plus score was used. The goal was to find full test administrations spanning the full NRS EFL range and beyond. In this study, portfolios were used with scale scores from 351 to 694. Judges were presented with the student portfolios in order from lowest to highest scale score.

6.1.ii. Participants

The 2012 standard-setting study panel included ten highly qualified judges representing wide geographic diversity. The following two tables provide information on the demographics of the judges. All ten of the judges were female. Table 49 shows the number of years judges had spent in an adult ESL career; half of them had more than 21 years of experience.

Table 50 shows the highest level of education attained by the judges: six had a master's degree and one had doctoral studies. Table 51 shows ethnicity of the judges, seven of whom were white, and one each was Asian, Hispanic, or mixed.

Table 49. Number of years experience judges had in adult ESL

Number of Years	Frequency
1-5 years	1
6-10 years	0
11-15 years	3
16-20 years	1
21 or more years	5
Total	**10**

Table 50. Highest education level attained by judges

Highest level of education	Frequency
Bachelor's degree	1
Some graduate study	2
Master's degree	6
Some doctoral study	1
Doctoral degree	0
Total	**10**

Table 51. Ethnicity of judges

Ethnicity	Frequency
White	7
Hispanic	1
Asian	1
Mixed	1
Total	**10**

Judges rated their familiarity with the NRS EFL descriptors and with BEST Plus on a 3-point scale: (3) Very Familiar, (2) Somewhat Familiar, or (1) Not Familiar. Table 52 presents an overview of these results. As the results suggest, the judges felt quite familiar with the NRS EFLs and even more familiar with BEST Plus.

Table 52. Judges' familiarity with the NRS EFLs and BEST Plus

	N	Minimum	Maximum	Mean
Familiarity with NRS	10	2	3	2.7
Familiarity with BEST Plus	10	3	3	3.0
Total	**10**	**2**	**3**	**2.85**

6.1.iii. Procedures

The study followed procedures consistent with a Body of Work procedure (Kingston, Kahl, Sweeney, & Bay, 2001), modified to make the procedure more efficient and reliable (see Kenyon & Fidelman, 2009). The procedure is the same as that which was used in the original 2002 study. This method was selected as most appropriate to make judgments based on videos of performances on the test. For example, the traditional Body of Work method generally uses two separate steps: a range-finding round and a pinpointing round. As Sweeney and Ferdous (2007) point out, doing two separate rounds requires a much larger number of portfolios than the number used in the BEST Plus standard-setting study described here.

Twenty-one videotaped full test administrations were carefully selected in advance to represent a variety of students demonstrating the entire range of performances and proficiency levels.

For the purposes of this section, each NRS EFL may be referred to interchangeably by either its full name or its level number. For example, the lowest NRS EFL, Beginning ESL Literacy, may be referred to as level 1; the Low Beginning ESL level may be referred to as level 2, and so forth. The names and corresponding numbers of the NRS EFLs are shown in Table 53.

Table 53. Revised NRS Educational Functioning Levels

NRS EFL	Descriptor
1	Beginning ESL Literacy
2	Low Beginning ESL
3	High Beginning ESL
4	Low Intermediate ESL
5	High Intermediate ESL
6	Advanced ESL
7	Exit of NRS EFLs

After viewing each student's full test administration, the judges applied three steps as they assigned student performances to the NRS EFLs. These are the three steps as presented in the instructions to the judges:

1. Decide at which NRS EFL (1 – 7) you feel the student in the videotaped test administration is currently functioning.
2. Think about how confident you are that this student is currently functioning at that NRS EFL (100% - 50%).
3. If not 100% confident of your selection in #2, decide at which adjacent NRS EFL (higher or lower) this student might also be functioning.

Thus, after watching an examinee's performance during a full test session, a judge in the NRS study could award, for example, 50% to level 2 and 50% to level 3, or 80% to level 2 and 20% to level 1, or 100% to level 3. All judgments are required to be in increments of 10% and may be centered on any one NRS EFL (1 through 6 and exit, also referred to as level 7) or on multiple adjacent levels.

After completing a practice round with three full test administrations as samples from the beginning, middle, and high end of the performance continuum, the judges began rating the 21 performances. First, all of the judges watched each videotaped full test administration once and marked their decisions on a scoring sheet during round one. When they all had finished, they shared their decisions, which were input into a table and shown immediately to the entire group projected on a screen in the room. The table also displayed the average rating in each category.

The judges then considered the outcome and discussed it as necessary. Judges whose ratings were outliers discussed their rationale for rating as they did. After this discussion, the judges did their second round of ratings. For the second round, judges watched the same videotaped full test administration again and then made their final ratings individually. The results from the second round were analyzed to determine the final cuts.

6.1.iv. Results

In order to determine the extent to which the judges agreed with each other across the two rounds of rating, a score was calculated for each judge for each profile. This was done by multiplying each NRS EFL by the percent confidence reported by the judge at that level during each round of rating. For example, the calculation for a judge indicating that they felt that a profile was 50% level 1 and 50% level 2 is shown in Equation 2.

Equation 2. Calculation of level judgment

$$1.5 = 0.5(1) + 0.5(2)$$

The results of these calculations for Round 1 can be seen in Figure 2. Figure 3 shows these calculations for Round 2.

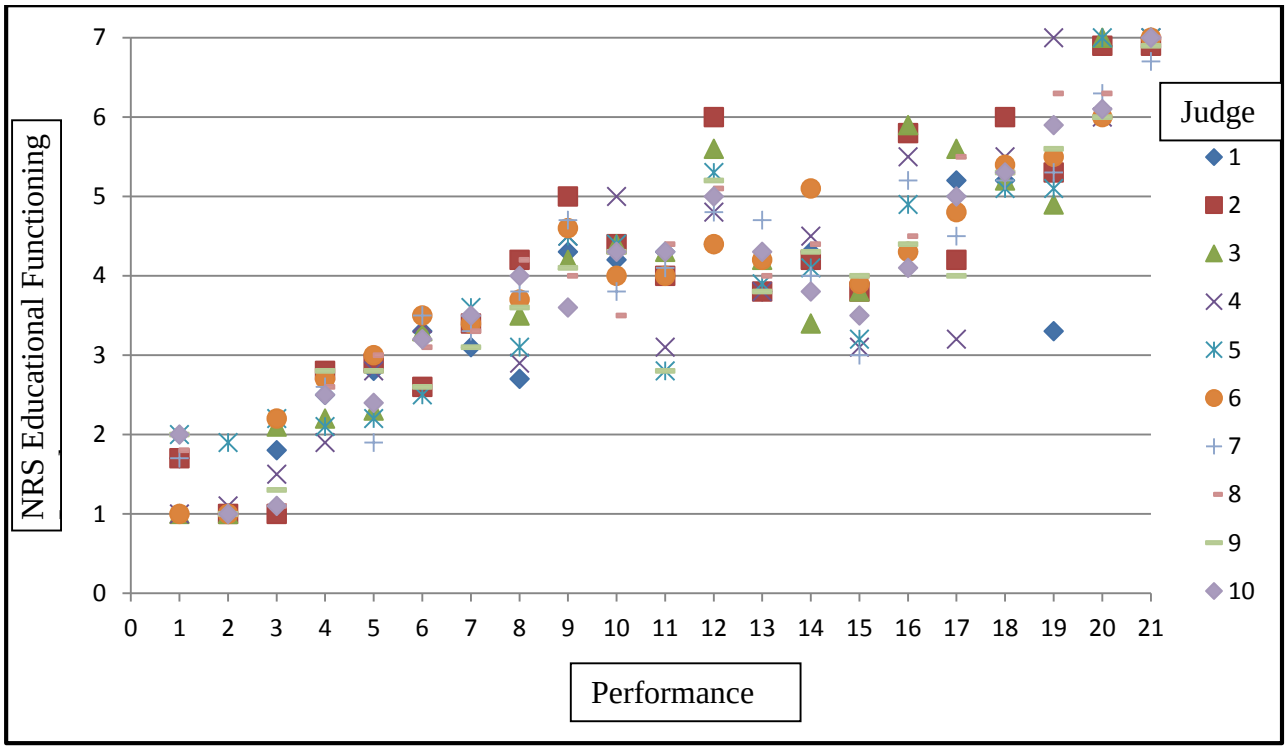

Figure 2. Judges' ratings during Round 1

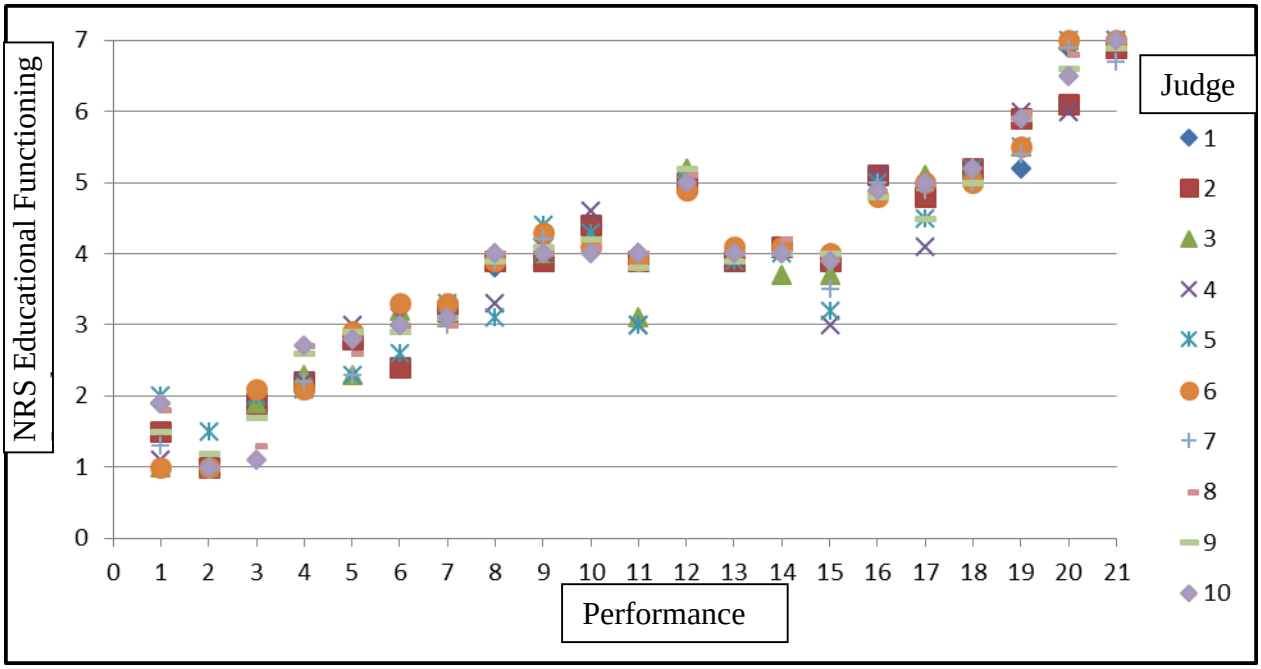

Figure 3. Judges' ratings during Round 2

As can be seen from Figure 2, during the first round of rating, judges' ratings ranged across levels, with some profiles garnering ratings across as many as 5 levels (e.g., Performance 19). During the second round of rating, as seen in Figure 3, judges' ratings for each profile generally fell within one NRS EFL of each other. For some profiles, the ratings fell in a very close range. For Profile 13, for example, the lowest rating was 3.9 and the highest was 4.1. The greatest disagreement can be seen in Profile 11, in which the highest weighted rating was 4 and the lowest was 3. Overall, the range of 21 profiles is within one level of difference during the second round of rating.

Table 54 shows the means and standard deviations of ratings from all judges for each profile across each round. Between Round 1 and Round 2 the standard deviations decreased, indicating that agreement increased.

Table 54. Means and standard deviations of two rounds of judging

Profile	Round 1 Mean	Round 1 Standard Deviation	Round 2 Mean	Round 2 Standard Deviation
1	1.59	0.43	1.50	0.39
2	1.10	0.28	1.07	0.16
3	1.54	0.49	1.70	0.39
4	2.47	0.31	2.33	0.24
5	2.61	0.38	2.68	0.28
6	3.10	0.39	2.95	0.27
7	3.36	0.17	3.14	0.11
8	3.57	0.53	3.78	0.32
9	4.35	0.40	4.11	0.16
10	4.23	0.40	4.20	0.19
11	3.81	0.65	3.67	0.44
12	5.06	0.50	5.04	0.10
13	4.05	0.30	3.97	0.07
14	4.21	0.45	4.02	0.13
15	3.62	0.39	3.72	0.37
16	4.89	0.67	4.93	0.13
17	4.69	0.74	4.78	0.32
18	5.35	0.25	5.09	0.10
19	5.42	0.97	5.68	0.29
20	6.45	0.45	6.68	0.37
21	6.95	0.10	6.95	0.10

Following the modified Body of Work method, logistic regression was used to determine, from the data collected from the judges, the point along the underlying proficiency continuum at which at least 50% of the judges would be expected to agree that the portfolio represents the work of the next higher proficiency level rather than the current proficiency level. Logistic regression is used when the outcome variable (dependent variable) is dichotomous. In this study, the dichotomy is either being rated at the lower proficiency level or the higher proficiency level. In other words, when conducting the analysis between two levels (e.g., 2 and 3), the input data were treated dichotomously as the percent at the lower level (2) or below and the percent at the higher level (3) or above. The BEST Plus scale score of the student who produced each portfolio was used as the indicator of student proficiency.

To illustrate how logistic regression was used, Table 55 shows an example of the data that were input to determine the cut score between proficiency levels 2 and 3 of the NRS EFLs. The first column shows the performance number, the second column shows the corresponding BEST Plus scale score, the third column shows the observed percent of weighting from the judges for whom that portfolio did not yet demonstrate ability at NRS EFL 3, and the fourth column shows the observed percent of weighting from the judges for whom that portfolio represented ability at least at level 3. These numbers represent an average across all 10 judges from the second round of rating.

Table 55. Example data for determination of NRS EFL 3 cut score

Portfolio Number	Scale Score	Percent Not Yet 3	Percent 3 or above
1	351	100	0
2	354	100	0
3	371	100	0
4	427	67	33
5	430	32	68
6	438	11	89
7	449	0	100
8	456	0	100
9	457	0	100
10	465	0	100
11	469	0	100
12	479	0	100
13	485	0	100
14	485	0	100
15	486	0	100
16	488	0	100
17	496	0	100
18	515	0	100
19	528	0	100
20	566	0	100
21	694	0	100

In this example, we see that until a score of 427 was reached, the judges were unanimous that the performance did not yet meet the criteria for NRS EFL 3. We also see that once a score of 449 was reached, judges were unanimous that the performance was at level 3 or above. Somewhere between those two scores is a point at which at least 50% of the weighting would be at least at level 3. From the observed data, that point most likely lies between portfolio 4 (total scale score of 427 with 33% agreement) and portfolio 5 (total scale score of 430 with 68% agreement).

Figure 4 is the graphic representation of Table 55, including the predicted logistic regression line. The vertical axis represents percentages. The horizontal axis represents the BEST Plus scale score. The 21 dots in Figure 4 represent the observed percentage of agreement among the judges that each of the 21 portfolios (each at its own scale score) represents a performance at level 3 or higher of the NRS EFLs. The curve in Figure 4 represents the predicted percentages fitting the logistic regression line to the data.

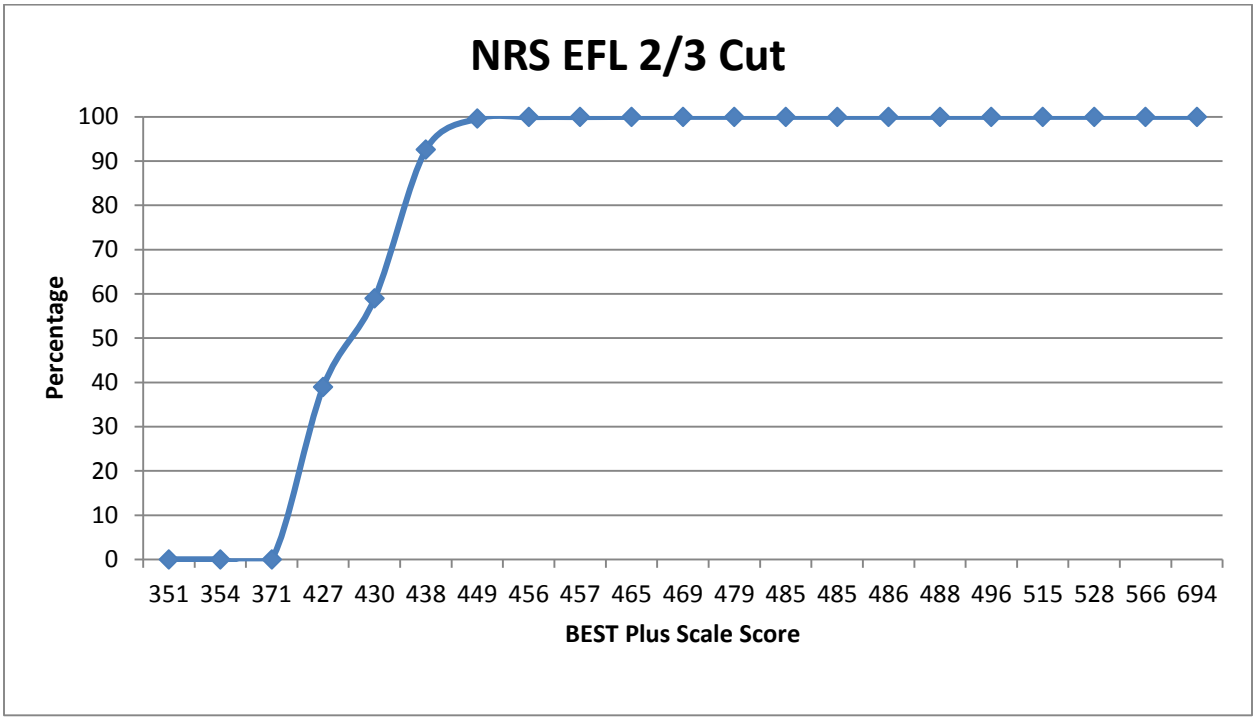

Figure 4. Example determination of a cut score

Visually, to find the point at which at least 50% of the weighting would be at level 3 using this figure, find 50 on the vertical axis, follow the horizontal line across to the point where it meets the curve, and go down to find the corresponding scale score on the horizontal axis. This scale score represents the cut between level 2 and level 3, because a group of judges (as represented by the judges in this study) would be more likely to rate a performance at that scale score at level 3 or above rather than at level 2 or below. In actuality, the exact point is found by solving a mathematical equation to determine the cut score, as illustrated in Figure 5.

Ln(Odds) = -115.9257 + (0.2704) x (scale score)

Figure 5. Paramater estimates for NRS EFL 2/3 cut

We are looking for the scale score at which the odds for a portfolio to be put in level 2 (or lower) or in level 3 (or higher) are equal. The odds at that scale score are 1 (1/1=1). Therefore, the left side of the equation is 0 as the natural logarithm of 1 is 0 as shown in Figure 6.

0 = -115.9257 + (0.2704) x (scale score)
115.9257 = 0.2704 x (scale score)
428.7193 = scale score

Figure 6. Calculation of NRS EFL 3 cut score

Because the results are typically not whole numbers, results were truncated to obtain a whole number scale score of the cut score set by the judges. In the above example, the number would be 428.

Table 56 shows the cut scores and standard errors that resulted from this standard-setting study. Beginning ESL Literacy is not included because all examinees falling below Low Beginning ESL are classified as Beginning ESL Literacy. The standard error at each cut score was calculated as the standard deviation of individual judge-based cut scores divided by the square root of the number of judges. For example, a total of ten judges participated in the current study. For the cut score at High Beginning ESL, a logistic regression analysis was conducted for each of the ten judges based on their individual second round ratings. This led to ten judge-based cut scores for High Beginning ESL. The standard error at High Beginning ESL was the standard deviation of the ten individual judge-based cut scores divided by the square root of 10.

Table 56. Cut scores from 2012 standard-setting study

	Low Beginning ESL	**High Beginning ESL**	**Low Intermediate ESL**	**High Intermediate ESL**	**Advanced ESL**	**Exit of NRS EFLs**
Cut score	362	428	453	485	525	565
Standard Error	3.39	1.28	1.59	0.64	1.28	10.71

Table 56 shows that for cut scores of High Beginning ESL through Advanced ESL, the standard errors were below 2, suggesting that these cut scores might be expected to be quite similar if they were determined based on similar procedures but a different panel of experts. Standard errors for cut scores at the two ends of any scale are expected to be larger, which was the case here for Low Beginning ESL and Exit of NRS EFLs. However, considering the wide range of the scale, from 88 to 999, the standard errors for cut scores at Low Beginning ESL and Exit of NRS EFLs were minor.

The 2002 standard-setting study translated scale scores on BEST Plus to Student Performance Levels (SPLs), which were commonly used in the field at the time. The SPLs were linked to the NRS EFLs in the then-current NRS EFL descriptors, which allowed for BEST Plus scale scores to be linked to NRS EFLs. The NRS EFLs were then modified in 2006, with new linkages to the

SPLs, allowing CAL to adjust the scale scores. Table 57 shows how those cut scores relate to the 2012 cut scores.

Table 57. 2006 and 2012 cut scores

	Low Beginning ESL	High Beginning ESL	Low Intermediate ESL	High Intermediate ESL	Advanced ESL	Exit of NRS EFLs
2006	401	418	439	473	507	541
2012	362	428	453	485	525	565

As Table 57 shows, the alignments of the new to the old scores are close, although the NRS EFLs do not align directly. Except for Low Beginning ESL, which has a lower cut score in 2012, all 2012 cut scores are 10 to 24 scale score points higher. Also, in all cases except Low Beginning ESL, the new cut was not placed beyond the original cut for the next highest level. For instance, the low intermediate ESL cut score was 439 in 2006, and slightly increased to 453 in 2012. This suggests consistency in the interpretation of the scale scores in relation to the NRS EFLs across the 2002 and 2012 standard-setting studies with different judges. In addition, it is a reminder that the older SPLs are only an approximation of NRS EFLs, but do not align perfectly with them, which is to be expected.

6.2. Adoption of 2012 cut scores

In September, 2013, CAL convened an External Advisory Board (EAB) of field experts to review the new cut scores and advise on policies for their use. The specific goals of the meeting were as follows:

- Review the 2012 BEST Plus NRS Standard-setting Study results
- Elicit recommendations regarding new BEST Plus cuts scores
- Elicit recommendations for implementation of the new cut scores

6.2.i. Participants

Seven specialists with deep familiarity with the BEST Plus exam, the field of adult ESL education, and the NRS attended the meeting. The following tables provide information on the demographics of the participants.

Table 58. Years experience participants had in adult ESL

Years	Count
1-5 years	0
6-10 years	3
11-15 years	1
16-20 years	0
21 or more years	3
Total	**7**

As Table 58 shows, over half of participants had more than 10 years experience in the field. Table 59 shows the highest level of education achieved by the participants.

Table 59. Highest education level achieved by participants

Highest education level	Count
Bachelor's degree	0
Some graduate study	1
Master's degree	4
Some doctoral study	1
Doctoral degree	1
Total	**7**

As Table 59 shows, all participants reported some education beyond a Bachelor's degree. Four had a Master's degree, one was working on a doctoral degree, and one had a doctorate. Table 60 shows the ethnicity of the participants.

Table 60. Ethnicity of participants

Ethnicity	Count
White	5
Hispanic	1
Asian	1
Mixed	0
Total	**7**

As Table 60 shows, five participants were white, one participant was Hispanic, and one was Asian.

In addition to collecting demographic information, participants were surveyed regarding their familiarity with NRS EFLs, SPLs, and BEST Plus. Table 61 shows how familiar the participants reported they were with the NRS EFL descriptors, and with BEST Plus. They rated their familiarity on a 3-point scale: (3) *Very Familiar*, (2) *Somewhat Familiar*, or (1) *Not Familiar*.

Table 61. Participants' familiarity with the NRS, BEST Plus, and SPLs

	N	Minimum	Maximum	Mean
Familiarity with NRS EFLs	7	2	3	2.7
Familiarity with SPLs	7	2	3	2.3
Familiarity with BEST Plus	7	3	3	3.0

As Table 61 shows, the EAB members felt quite familiar with the NRS EFLs and even more familiar with BEST Plus. They were less familiar with the SPLs, but that is not considered problematic as the standard study aligned the scale scores directly with the NRS EFLs.

6.2.ii. Procedures

The meeting started with a description of the cut-score study of 2012, why it was conducted, how it was conducted, and who participated in the study. The EAB then participated in a sample cut score activity using two samples that were rated in the 2012 standard-setting study. Similar to the 2012 study, the participants were asked to identify the NRS EFLs of both of the students and to pinpoint their percentage of confidence in their judgment, as well as the other level (either above or below the selected level) the performance might be. After the discussion of the scores, the results from the standard-setting study were displayed. The participants' ratings were consistent with the levels recommended by the standard-setting study in 2012. The confidence in these levels was high.

Next, CAL presented the changes to BEST Plus cut scores based on the September, 2012 cut score study. Participants were asked to discuss privately (without CAL staff in the room) what their recommendations should be for the new cut scores for the NRS EFLs. The participants were asked to vote, in a private written ballot, on acceptance of the new cut scores and on the timing of their release. Then, after further discussion of the issue, the participants voted again, this time including their names on the ballot. Finally, participants were asked to rate their confidence in the new cut scores to adequately and appropriately classify students in relation to the NRS EFLs.

6.2.iii. Results

Participants unanimously decided to accept the 2012 cut scores and for them to go into effect July 1, 2015 (some 22 months hence). The reasoning for the delay was that the impact of the change would be far-reaching and many new materials would need to be developed and distributed. Participants also reported high confidence in the cut scores recommended by the EAB to adequately and appropriately classify students in relation to the NRS EFLs. The rating options were on a four-point scale with *Very high* (4), *High* (3), *Medium* (2), *Low* (1), and *N/A* (0). Means were calculated. Table 62 shows that the confidence was strong in all the levels, ranging from means of 3.1 to 3.5; the 3.1 mean reflected slightly less confidence in the alignment for moving from Beginning ESL Literacy to Low Beginning ESL and from Advanced ESL to Exit of NRS EFLs.

Table 62. Participants' rate of confidence in alignment of cut score with NRS EFLs

	Beginning ESL Literacy/ Low Beginning ESL (1/2)	**Low Beginning ESL/High Beginning ESL (2/3)**	**High Beginning ESL/Low Intermediate ESL (3/4)**	**Low Intermediate ESL/High Intermediate ESL (4/5)**	**High Intermediate ESL/Advanced ESL (5/6)**	**Advanced ESL/Exit NRS (6/Exit NRS)**
Mean	3.1	3.5	3.5	3.5	3.5	3.1

6.2.iv. Rationale for using 2012 results

As a result of the 2012 standard-setting study and 2013 review of the results by the External Advisory Board (EAB), CAL decided to adopt the new cut scores effective July 1, 2015. These new cut scores will remain in effect until the NRS EFLs are next revised. This decision is supported by a number of factors.

1. **Use of current framework.** The 2012 standard-setting study was undertaken with the intention of relating performances on BEST Plus scale scores directly to the current National Reporting System (NRS) Educational Functioning Levels (EFLs), which were revised in 2006. The current cut scores resulted from a 2002 standard-setting study which translated scale scores into Student Performance Levels (SPLs) which were commonly used at the time to report student proficiency. Because the relationship between SPLs and NRS EFLs was established, scale scores were connected to the NRS EFLs via the SPLs. The 2012 study provides a direct interpretation of scale scores to NRS EFLs, without dependence on the older SPLs.
2. **Comparability to current cut scores**. The 2012 cut scores are not drastically different from the current cut scores, indicating that the new scores provide refinements to the classifications, not a paradigm shift.
3. **Level of agreement**. As indicated in section 6.1.iv, the ten judges who participated in the standard-setting study had a high level of agreement, with no disagreement on any one profile exceeding one level.
4. **EAB acceptance**. The External Advisory Board (EAB), which met in September, 2013 to review the results of the 2012 standard-setting study, voted unanimously to accept the new cut scores effective July 1, 2015. Furthermore, they expressed high levels of satisfaction with the process and with the new cut scores themselves.

7. References

Agresti, A. (1996). *An introduction to categorical data analysis* (Vol. 135). New York, NY: Wiley.

Center for Applied Linguistics. (2005). *BEST Plus Technical Report.* Washington, DC: Author.

Chi, Y., Garcia, R.B., Surber, C., & Trautman, L. (2011). Alignment study between the common core state standards in English language arts and mathematics and the WIDA English language proficiency standards, 2007 edition, prekindergarten through grade 12. University of Oklahoma College of Continuing Education eteam.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of tem exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment 5*(8). Retrieved December 7, 2014 from http://www.jtla.org.

Goldstein, H. (1983). Measuring changes in educational attainment over time: problems and possibilities. Journal of Educational Measurement, 20(4), 369–377.

Huynh, H. & Meyer, P. (2010). Use of Robust z in detecting unstable items in item response theory models. Practical Assessment, Research & Evaluation, 15(2). Available online: http://pareonline.net/getvn.asp?v=15&n=2.

Kenyon, D. & Fidelman, C. (2009, April). Standard Setting with the Modified Body of Work Method. Paper presented at National Council on Measurement in Education, San Diego.

Kingston, N.M., Kahl, S.R., Sweeney, K.P., Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* . Mahwah, NJ: Lawrence Erlbaum Associates.

Kim, D., Barton, K., & Choi, S. (2010, May). Sample size impact on screening methods in the Rasch model. Paper presented at American Educational Research Association, Denver.

Linacre, J. M. (2012). *Winsteps Rasch Tutorials*. Retrieved from http://www.winsteps.com/a/winsteps-tutorial-3.pdf.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32,* 179-197.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420–428. doi:10.1037/0033-2909.86.2.420

Sweeney, K. P., & Ferdous, A. (2007, April). *Variations of the "Body of Work" standard setting method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Young, M. J., & Yoon, B. (1998, April). Estimating the consistency and accuracy of classifications in a standards-referenced assessment (CSE Tech. Rep. No. 475). Los Angeles: University of California, Los Angeles; National Center for Research on Evaluation, Standards, and Student Testing.

Young, S. (2007). Effects of instructional hours and intensity of instruction on NRS level gain in listening and speaking. CAL Digest. Washington, DC: Center for Applied Linguistics.

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
warmup.2.WA	WA	-2.33	245	401	155	189	146	89	128	1353	100%
warmup.3.WA	WA	-1.42	245	401	155	189	146	89	128	1353	100%
warmup.4.WA	WA	-1.48	245	401	155	189	146	89	128	1353	100%
warmup.5.WA	WA	0.51	245	401	155	189	146	89	128	1353	100%
warmup.6.WA	WA	-2.11	245	401	155	189	146	89	128	1353	100%
warmup.7.WA	WA	0.19	245	401	155	189	146	89	128	1353	100%
hous.2.eq	EQ	-0.96	50	46	22	32	42	32	25	249	18%
educ.1.EQ	EQ	-0.10	35	74	34	33	28	20	23	247	18%
fami.2.EQ	EQ	-1.60	31	60	16	43	35	27	27	239	18%
*old.comm.3.EQ	EQ	-0.50	34	58	22	39	30	22	25	230	17%
heal.2.EQ	EQ	-1.54	55	43	22	23	29	14	35	221	16%
*old.recr.1.EQ	EQ	-0.57	9	60	26	38	29	18	36	216	16%
*old.fami.3.PD	PD	-1.28	62	39	16	24	36	17	22	216	16%
cons.2.EQ	EQ	-0.70	50	39	8	29	34	20	26	206	15%
heal.1.EQ	EQ	-0.45	14	58	22	32	28	15	35	204	15%
tran.1.EQ	EQ	-0.14	35	52	27	27	12	23	26	202	15%
*old.cons.3.EQ	EQ	-0.75	22	59	20	25	32	24	18	200	15%
cons.1.EQ	EQ	-0.57	26	53	26	28	12	14	25	184	14%
comm.2.EQ	EQ	-0.31	41	48	20	15	12	16	31	183	14%
weat.2.PD	PD	-1.05	36	58	20	37	21	3	3	178	13%
*old.comm.1.EQ	EQ	0.25	8	28	25	33	38	23	21	176	13%
weat.2.GE	GE	-0.11	23	37	19	40	35	14	7	175	13%
educ.3.EQ	EQ	-0.16	26	33	22	26	18	17	32	174	13%
*old.recr.2.EQ	EQ	-0.81	21	46	21	32	15	19	20	174	13%
*old.heal.3.EQ	EQ	0.48	32	52	17	16	16	17	21	171	13%
tran.2.EQ	EQ	-0.72	29	34	11	13	26	20	34	167	12%
*old.civi.1.EQ	EQ	-0.38	3	28	22	35	26	14	35	163	12%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
heal.4.EQ	EQ	-0.75	21	43	21	18	20	18	21	162	12%
*old.weat.1.PE	PE	0.26	13	37	23	27	28	24	7	159	12%
recr.3.PE	PE	-0.64	36	52	23	33	8	3	2	157	12%
recr.3.PD	PD	-1.73	76	56	10	5	5	2	2	156	12%
recr.3.EQ	EQ	0.34	14	12	18	36	40	16	14	150	11%
educ.1.PD	PD	-1.39	41	70	19	7	4	3	6	150	11%
tran.3.PE	PE	0.05	1	31	20	35	38	15	7	147	11%
*old.educ.2.EQ	EQ	-0.11	35	33	17	11	5	17	28	146	11%
heal.2.PD	PD	-1.57	72	41	10	8	5	4	5	145	11%
educ.1.CQ	CQ	-1.20	9	68	34	26	7	0	0	144	11%
*old.hous.2.PE	PE	0.06	0	12	18	32	40	29	8	139	10%
*old.hous.2.PD	PD	-1.78	68	42	10	5	3	5	5	138	10%
*old.weat.1.PD	PD	-0.99	23	50	31	24	5	2	3	138	10%
*old.heal.3.PD	PD	-1.09	32	53	17	15	8	6	5	136	10%
*old.hous.1.EQ	EQ	-0.63	37	43	15	10	6	4	20	135	10%
*old.weat.1.EQ	EQ	0.62	10	13	8	7	27	24	41	130	10%
fami.2.PD	PD	-1.61	31	59	12	16	6	3	3	130	10%
comm.2.PD	PD	-1.40	51	46	10	7	7	2	3	126	9%
weat.2.EQ	EQ	1.16	11	22	5	10	19	22	36	125	9%
cons.2.PD	PD	-1.34	72	37	2	3	2	3	4	123	9%
weat.2.YN	YN	-0.88	11	43	17	28	18	4	1	122	9%
*old.hous.1.PD	PD	-1.56	43	43	15	10	5	2	3	121	9%
tran.1.PD	PD	-1.35	36	52	15	9	3	4	2	121	9%
on-t.3.EQ	EQ	-0.30	10	17	15	22	16	15	25	120	9%
heal.2.CQ	CQ	-1.20	24	43	22	18	11	1	0	119	9%
*old.comm.3.PD	PD	-1.47	34	55	13	11	0	2	3	118	9%
*old.hous.2.YN	YN	-1.82	29	46	21	14	4	2	0	116	9%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
*old.recr.1.CQ	CQ	-0.54	4	22	23	37	23	4	0	113	8%
tran.1.CQ	CQ	-0.84	14	39	27	26	6	1	0	113	8%
heal.1.PD	PD	-1.21	14	55	15	12	6	7	3	112	8%
cons.1.CQ	CQ	-1.01	6	43	26	26	8	1	1	111	8%
cons.1.PD	PD	-1.51	26	53	8	10	6	3	5	111	8%
*old.weat.1.YN	YN	-0.97	10	33	31	27	8	1	1	111	8%
*old.cons.3.PD	PD	-1.34	22	57	10	6	6	5	4	110	8%
fami.2.PE	PE	0.27	0	6	13	35	33	17	5	109	8%
recr.3.EL	EL	0.36	0	2	15	30	40	13	8	108	8%
tran.2.PD	PD	-1.31	44	34	10	9	6	2	3	108	8%
*old.hous.3.EQ	EQ	1.12	11	11	3	10	14	20	37	106	8%
*old.hous.3.PD	PD	-1.35	11	34	16	24	14	2	5	106	8%
comm.2.YN	YN	-1.02	15	46	20	15	7	2	1	106	8%
fami.2.CQ	CQ	-1.03	14	44	18	23	6	0	0	105	8%
educ.1.PE	PE	0.23	0	5	18	32	27	15	7	104	8%
*old.civi.3.PD	PD	-1.22	26	1	5	14	29	11	17	103	8%
*old.fami.3.EQ	EQ	-1.65	45	39	7	4	2	2	3	102	8%
*old.recr.1.PD	PD	-1.33	9	57	14	12	7	1	2	102	8%
*old.recr.2.PD	PD	-1.37	21	43	12	12	8	3	2	101	7%
*old.hous.3.PE	PE	0.47	0	23	11	20	17	14	16	101	7%
tran.3.EQ	EQ	0.90	0	0	3	18	35	18	25	99	7%
gett.3.EQ	EQ	-0.16	9	16	9	24	15	13	12	98	7%
on-t.2.EQ	EQ	-0.93	16	24	8	13	10	8	19	98	7%
heal.2.GE	GE	0.27	0	8	18	21	29	13	9	98	7%
heal.1.YN	YN	-0.70	10	30	22	26	8	1	0	97	7%
*old.on-t.1.PD	PD	-1.02	16	23	11	13	16	10	7	96	7%
cons.2.EL	EL	0.16	0	3	7	29	34	12	10	95	7%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
heal.4.PD	PD	-1.24	21	41	13	7	7	2	4	95	7%
recr.3.CQ	CQ	-1.57	35	53	4	1	0	0	0	93	7%
*old.heal.3.CQ	CQ	-1.14	11	49	17	12	3	0	0	92	7%
tran.3.PD	PD	-1.34	1	31	21	23	7	3	6	92	7%
*old.cons.3.CQ	CQ	-0.96	11	36	13	17	11	3	0	91	7%
gett.1.EQ	EQ	-0.28	8	15	10	14	16	11	16	90	7%
*old.fami.3.PE	PE	0.33	0	2	10	24	36	12	6	90	7%
*old.comm.3.EL	EL	0.10	0	3	4	33	29	12	7	88	7%
civi.2.EQ	EQ	-0.27	14	23	8	8	8	4	23	88	7%
*old.gett.2.EQ	EQ	-0.18	4	24	10	14	12	10	14	88	7%
*old.comm.3.YN	YN	-0.95	12	42	21	11	1	0	0	87	6%
*old.recr.1.YN	YN	-0.85	9	55	11	8	3	1	0	87	6%
*old.educ.2.PD	PD	-1.53	37	31	7	2	4	3	2	86	6%
*old.recr.1.PE	PE	0.43	0	3	10	25	23	16	8	85	6%
*old.recr.2.CQ	CQ	-0.48	12	13	16	29	7	4	3	84	6%
cons.2.PE	PE	0.67	0	3	6	25	23	16	11	84	6%
*old.hous.1.CQ	CQ	-1.46	17	42	13	8	2	1	0	83	6%
heal.4.CQ	CQ	-1.11	13	34	21	12	1	1	0	82	6%
*old.comm.1.CQ	CQ	-0.38	0	0	16	33	29	3	0	81	6%
*old.hous.2.EL	EL	1.19	0	0	1	14	23	23	20	81	6%
*old.civi.1.PE	PE	1.11	0	1	9	26	16	13	16	81	6%
*old.fami.3.EL	EL	0.43	0	1	8	22	25	13	11	80	6%
educ.3.PE	PE	0.78	0	5	10	17	17	16	15	80	6%
*old.civi.1.YN	YN	-0.46	0	2	10	33	26	7	2	80	6%
tran.1.YN	YN	-1.34	36	30	7	5	1	1	0	80	6%
*old.comm.1.PD	PD	-0.97	8	30	23	11	3	2	2	79	6%
comm.2.CQ	CQ	-1.63	47	18	6	3	3	1	0	78	6%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
*old.cons.3.GE	GE	-0.58	0	12	19	24	18	5	0	78	6%
*old.comm.3.PE	PE	0.81	0	3	8	25	16	14	12	78	6%
*old.recr.2.YN	YN	-0.85	21	43	8	4	1	0	0	77	6%
heal.1.CQ	CQ	-0.97	14	49	6	5	1	0	0	75	6%
educ.3.YN	YN	-0.73	0	14	21	25	13	2	0	75	6%
*old.comm.3.CQ	CQ	-1.25	34	36	1	3	0	0	0	74	5%
educ.1.YN	YN	-1.44	39	29	5	1	0	0	0	74	5%
educ.3.PD	PD	-1.39	26	28	9	2	1	2	5	73	5%
heal.1.PE	PE	0.01	0	4	9	25	25	6	4	73	5%
fami.2.YN	YN	-1.22	30	36	2	4	1	0	0	73	5%
*old.hous.2.CQ	CQ	-2.31	51	19	2	0	0	0	0	72	5%
fami.2.GE	GE	0.53	0	1	2	15	20	18	16	72	5%
*old.recr.2.PE	PE	0.31	0	3	9	24	14	17	5	72	5%
heal.2.YN	YN	-2.23	61	11	0	0	0	0	0	72	5%
weat.2.CQ	CQ	-1.34	36	34	1	0	0	0	0	71	5%
*old.hous.1.YN	YN	-1.52	38	13	9	8	1	1	0	70	5%
*old.hous.3.YN	YN	-0.49	11	34	15	8	2	0	0	70	5%
on-t.3.YN	YN	-0.75	9	17	15	17	7	3	2	70	5%
tran.1.PE	PE	0.74	0	0	9	18	10	22	9	68	5%
on-t.2.PD	PD	-1.16	23	24	7	2	4	2	5	67	5%
*old.comm.1.PE	PE	0.70	0	0	2	17	23	17	8	67	5%
*old.heal.3.YN	YN	-1.48	32	22	7	4	2	0	0	67	5%
*old.hous.3.CQ	CQ	-0.24	11	11	3	23	15	2	0	65	5%
tran.2.EL	EL	0.57	0	0	3	11	26	17	8	65	5%
fami.1.EQ	EQ	-0.01	0	0	3	9	18	10	24	64	5%
on-t.3.PD	PD	-1.49	11	31	5	6	3	5	3	64	5%
on-t.3.PE	PE	-0.54	3	21	12	16	3	5	3	63	5%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
*old.civi.1.PD	PD	-1.31	3	28	11	3	6	4	7	62	5%
gett.3.YN	YN	-0.68	3	13	8	24	12	2	0	62	5%
civi.2.CQ	CQ	-0.64	14	24	8	8	4	1	2	61	5%
cons.2.CQ	CQ	-1.47	22	38	1	0	0	0	0	61	5%
cons.1.YN	YN	-1.45	26	28	1	4	1	0	1	61	5%
*old.weat.1.CQ	CQ	-1.01	23	30	5	1	1	0	0	60	4%
cons.1.GE	GE	1.18	0	0	14	17	6	10	13	60	4%
tran.3.CQ	CQ	-0.47	0	0	9	32	14	4	0	59	4%
comm.2.PE	PE	0.95	0	2	11	9	8	15	14	59	4%
recr.3.YN	YN	-2.57	55	4	0	0	0	0	0	59	4%
*old.fami.3.CQ	CQ	-1.44	12	38	8	0	0	0	0	58	4%
*old.cons.3.PE	PE	0.50	0	0	2	4	21	21	10	58	4%
*old.cons.3.YN	YN	-1.13	22	35	1	0	0	0	0	58	4%
*old.educ.2.CQ	CQ	-1.10	1	24	18	10	2	1	1	57	4%
tran.2.CQ	CQ	-1.41	5	34	10	6	1	1	0	57	4%
civi.2.PD	PD	-1.60	14	24	7	2	4	1	4	56	4%
heal.4.PE	PE	0.71	0	2	11	14	14	9	6	56	4%
weat.2.PE	PE	0.70	0	0	2	3	7	16	28	56	4%
heal.4.YN	YN	-1.13	21	23	5	4	1	1	0	55	4%
*old.on-t.1.EQ	EQ	-1.03	14	23	10	3	2	1	1	54	4%
tran.2.GE	GE	0.83	0	0	1	4	17	9	23	54	4%
*old.educ.2.PE	PE	0.91	0	2	11	10	4	15	12	54	4%
*old.civi.1.CQ	CQ	-0.63	3	28	14	4	3	1	0	53	4%
on-t.3.GE	GE	0.19	0	0	1	10	15	14	13	53	4%
*old.comm.1.YN	YN	-0.76	8	30	10	2	2	0	0	52	4%
on-t.2.CQ	CQ	-1.00	5	23	8	10	3	2	0	51	4%
educ.3.CQ	CQ	-1.19	26	19	2	2	0	1	0	50	4%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
gett.3.GE	GE	0.60	0	0	4	14	14	12	6	50	4%
heal.1.GE	GE	0.51	0	0	0	1	9	13	27	50	4%
cons.2.YN	YN	-2.20	50	0	0	0	0	0	0	50	4%
tran.3.YN	YN	-0.56	1	31	13	5	0	0	0	50	4%
*old.hous.3.EL	EL	0.40	0	0	1	3	9	15	21	49	4%
gett.1.GE	GE	0.74	0	2	4	11	14	11	7	49	4%
*old.weat.1.GE	GE	1.23	0	0	0	0	9	11	29	49	4%
civi.2.YN	YN	-0.77	14	23	7	2	2	1	0	49	4%
*old.fami.3.YN	YN	-1.86	49	0	0	0	0	0	0	49	4%
heal.4.EL	EL	0.64	0	0	0	6	14	15	13	48	4%
*old.educ.2.YN	YN	-1.46	36	10	0	0	1	1	0	48	4%
recr.3.GE	GE	0.60	0	0	0	4	17	11	14	46	3%
gett.3.PD	PD	-1.13	9	16	4	8	3	2	3	45	3%
on-t.3.CQ	CQ	-0.85	12	28	4	0	0	0	0	44	3%
*old.heal.3.PE	PE	0.77	0	1	0	4	14	17	8	44	3%
*old.comm.1.EL	EL	0.67	0	0	1	2	9	20	11	43	3%
*old.heal.3.EL	EL	0.84	0	0	0	5	13	14	11	43	3%
*old.gett.2.CQ	CQ	-0.56	4	24	10	4	0	0	0	42	3%
tran.2.YN	YN	-1.71	40	0	1	1	0	0	0	42	3%
heal.2.PE	PE	0.68	0	0	2	4	6	4	25	41	3%
on-t.2.YN	YN	-1.54	23	9	5	1	2	0	0	40	3%
*old.on-t.1.CQ	CQ	-1.13	1	22	11	4	0	1	0	39	3%
*old.civi.3.PE	PE	0.17	0	0	0	1	27	6	5	39	3%
gett.1.YN	YN	-0.83	0	8	10	14	7	0	0	39	3%
*old.gett.2.GE	GE	0.51	0	1	6	10	9	8	3	37	3%
tran.1.EL	EL	0.84	0	0	0	2	6	14	14	36	3%
educ.1.GE	GE	1.04	0	0	0	2	8	10	16	36	3%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
*old.gett.2.PD	PD	-1.33	4	23	4	3	1	0	1	36	3%
cons.1.EL	EL	0.90	0	0	0	4	8	12	11	35	3%
*old.civi.3.EL	EL	0.74	0	0	0	2	11	7	15	35	3%
on-t.2.PE	PE	0.83	0	0	1	12	8	3	11	35	3%
*old.gett.2.YN	YN	-0.22	4	6	3	10	10	1	1	35	3%
educ.3.EL	EL	0.96	0	0	0	1	5	8	20	34	3%
gett.1.PD	PD	-1.44	8	13	5	1	4	1	2	34	3%
tran.3.GE	GE	0.87	0	0	0	0	9	7	17	33	2%
*old.educ.2.EL	EL	0.96	0	0	0	1	3	8	20	32	2%
*old.cons.3.EL	EL	1.02	0	0	0	3	5	6	17	31	2%
on-t.2.EL	EL	0.69	0	0	1	3	7	8	12	31	2%
fami.1.GE	GE	0.17	0	0	0	5	16	8	2	31	2%
*old.on-t.1.GE	GE	0.54	0	0	0	8	12	8	3	31	2%
on-t.1.PE	PE	0.28	0	0	2	10	13	5	1	31	2%
comm.2.EL	EL	1.12	0	0	0	1	5	10	13	29	2%
weat.2.EL	EL	1.75	0	0	0	0	1	2	26	29	2%
*old.recr.2.GE	GE	1.09	0	0	0	2	4	10	13	29	2%
gett.3.CQ	CQ	-1.17	9	12	3	3	1	0	0	28	2%
heal.1.EL	EL	1.75	0	0	0	0	0	2	26	28	2%
heal.2.EL	EL	1.70	0	0	0	0	2	5	21	28	2%
*old.civi.3.YN	YN	-1.91	26	0	1	0	0	0	0	27	2%
fami.2.EL	EL	1.04	0	0	0	0	1	5	20	26	2%
tran.2.PE	PE	1.51	0	0	0	3	5	1	17	26	2%
cons.2.GE	GE	1.20	0	0	0	0	1	4	20	25	2%
civi.1.EL	EL	1.63	0	0	0	0	0	6	18	24	2%
on-t.3.EL	EL	1.75	0	0	0	0	0	5	19	24	2%
weat.1.EL	EL	1.76	0	0	0	0	1	0	23	24	2%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
comm.3.GE	GE	0.83	0	0	0	0	0	3	21	24	2%
heal.4.GE	GE	1.10	0	0	0	0	6	0	18	24	2%
hous.3.GE	GE	2.01	0	0	0	0	0	0	24	24	2%
recr.1.GE	GE	1.27	0	0	0	1	0	4	19	24	2%
*old.civi.3.CQ	CQ	-0.49	0	1	5	14	2	0	1	23	2%
fami.1.EL	EL	0.82	0	0	0	0	3	3	17	23	2%
*old.hous.1.EL	EL	1.38	0	0	2	2	5	3	11	23	2%
tran.3.EL	EL	1.24	0	0	0	0	2	3	18	23	2%
fami.1.PE	PE	1.31	0	0	0	0	2	2	19	23	2%
recr.1.EL	EL	1.54	0	0	0	0	0	1	20	21	2%
civi.2.EL	EL	1.06	0	0	0	0	3	3	13	19	1%
gett.1.EL	EL	0.92	0	0	0	0	5	4	10	19	1%
civi.2.GE	GE	1.07	0	0	0	0	4	3	12	19	1%
fami.3.GE	GE	1.11	0	0	0	0	0	1	18	19	1%
cons.1.PE	PE	1.24	0	0	0	1	1	2	15	19	1%
hous.1.PE	PE	1.68	0	0	2	2	3	1	11	19	1%
civi.1.GE	GE	1.73	0	0	0	0	0	0	18	18	1%
comm.2.GE	GE	1.37	0	0	0	0	1	0	17	18	1%
gett.1.CQ	CQ	-1.30	8	7	2	0	0	0	0	17	1%
gett.2.EL	EL	0.80	0	0	0	0	2	4	11	17	1%
gett.3.EL	EL	1.32	0	0	0	0	1	6	10	17	1%
comm.1.GE	GE	1.14	0	0	0	0	1	1	14	16	1%
educ.3.GE	GE	1.40	0	0	0	0	0	0	16	16	1%
on-t.1.YN	YN	-1.59	15	1	0	0	0	0	0	16	1%
educ.1.EL	EL	1.28	0	0	0	0	0	4	11	15	1%
*old.civi.3.EQ	EQ	-1.22	0	0	1	3	6	0	5	15	1%
educ.2.GE	GE	1.56	0	0	0	0	0	1	13	14	1%

Appendix A: Item-level exposure rates by NRS EFLs

Item Name	Item Type	Selection Difficulty	NRS EFL 1	NRS EFL 2	NRS EFL 3	NRS EFL 4	NRS EFL 5	NRS EFL 6	NRS EFL 7	Exposure Counts per Item	Exposure Rate
hous.1.GE	GE	1.65	0	0	0	0	1	2	11	14	1%
hous.2.GE	GE	1.38	0	0	0	0	1	3	10	14	1%
tran.1.GE	GE	1.38	0	0	0	0	1	0	13	14	1%
on-t.1.EL	EL	0.75	0	0	0	0	3	3	7	13	1%
recr.2.EL	EL	1.30	0	0	0	0	0	2	11	13	1%
fami.1.PD	PD	-1.42	0	1	1	3	3	1	4	13	1%
gett.1.PE	PE	0.80	0	0	0	0	4	1	8	13	1%
civi.3.GE	GE	1.01	0	0	0	0	0	2	10	12	1%
heal.3.GE	GE	1.43	0	0	0	0	0	0	12	12	1%
civi.2.PE	PE	1.22	0	0	0	0	2	0	10	12	1%
fami.1.CQ	CQ	-0.50	0	0	2	6	2	0	0	10	1%
on-t.2.GE	GE	1.04	0	0	0	0	0	0	10	10	1%
gett.2.PE	PE	1.16	0	0	0	1	0	1	8	10	1%
gett.3.PE	PE	1.23	0	0	0	0	0	2	6	8	1%
fami.1.YN	YN	-0.58	0	1	2	1	0	0	0	4	0%

*Items to be retired after the update of BEST Plus item bank (2.0)

Appendix B: Displacement values and Robust Z statistics of the 168 continuing items

Count	Item	Displacement	Robust Z
0-99	civi.1.EL	0.36	-2.20
0-99	civi.1.GE	0.32	-2.59
0-99	civi.2.EL	0.14	-1.29
0-99	civi.2.GE	-0.01	-0.89
0-99	civi.2.PE	-0.08	-0.81
0-99	civi.3.GE	-0.14	-0.62
0-99	comm.1.GE	0.48	-2.20
0-99	comm.2.EL	-0.17	-0.54
0-99	comm.2.GE	-0.16	-0.18
0-99	comm.3.GE	0.21	-1.51
0-99	cons.1.PE	0.07	-1.22
0-99	cons.2.GE	-0.37	0.60
0-99	educ.1.EL	-0.37	0.44
0-99	educ.2.GE	-0.50	0.72
0-99	educ.3.GE	-0.04	-0.66
0-99	fami.1.EL	-0.46	0.18
0-99	fami.1.PD	0.19	-0.36
0-99	fami.1.PE	0.39	-1.39
0-99	fami.1.YN	-0.27	1.34
0-99	fami.2.EL	-0.14	-0.66
0-99	fami.3.GE	-0.29	-0.24
0-99	gett.1.CQ	-0.40	1.37
0-99	gett.1.EL	0.15	-1.22
0-99	gett.1.PE	-0.24	0.54
0-99	gett.2.EL	-0.06	-0.72
0-99	gett.2.PE	0.34	-1.22
0-99	gett.3.CQ	-0.08	1.02
0-99	gett.3.EL	0.06	-1.28
0-99	gett.3.PE	0.27	-1.08
0-99	heal.1.EL	-0.32	-0.08
0-99	heal.2.EL	-0.14	-0.47
0-99	heal.3.GE	-0.35	0.23
0-99	heal.4.GE	0.01	-0.41
0-99	hous.1.GE	-0.33	-0.23
0-99	hous.1.PE	-0.16	-0.20
0-99	hous.2.GE	0.05	-1.31
0-99	hous.3.GE	-0.78	1.10
0-99	on-t.1.EL	0.05	-0.29
0-99	on-t.1.PE	-0.10	-0.05
0-99	on-t.1.YN	0.33	0.56

Appendix B: Displacement values and Robust Z statistics of the 168 continuing items

Count	Item	Displacement	Robust Z
0-99	on-t.2.GE	-0.25	0.57
0-99	on-t.3.CQ	-0.25	1.39
0-99	on-t.3.EL	-0.24	-0.36
0-99	recr.1.EL	-0.18	-0.36
0-99	recr.1.GE	0.05	-1.07
0-99	recr.2.EL	1.14	-3.10
0-99	tran.1.GE	0.00	-1.26
0-99	tran.2.PE	0.25	-1.61
0-99	tran.3.EL	0.07	-0.75
0-99	weat.1.EL	-0.54	0.51
0-99	weat.2.EL	-0.09	-0.81
100-199	civi.2.YN	-0.19	1.16
100-199	comm.2.PE	0.35	-0.63
100-199	cons.1.EL	-0.04	-0.59
100-199	cons.1.GE	0.15	-0.86
100-199	cons.2.CQ	-0.10	0.92
100-199	cons.2.YN	0.44	0.24
100-199	educ.1.GE	0.11	-0.47
100-199	educ.3.CQ	0.07	0.95
100-199	educ.3.EL	0.19	-1.31
100-199	fami.1.CQ	0.05	0.48
100-199	fami.1.GE	0.12	-0.48
100-199	gett.1.GE	0.18	-0.90
100-199	gett.1.PD	0.52	-0.74
100-199	gett.1.YN	-0.30	1.07
100-199	gett.3.GE	0.11	-0.24
100-199	gett.3.PD	0.00	-0.08
100-199	heal.1.GE	-0.24	0.00
100-199	heal.2.PE	0.03	0.00
100-199	heal.2.YN	-0.16	0.87
100-199	heal.4.EL	0.22	-1.14
100-199	heal.4.YN	-0.30	1.33
100-199	on-t.2.CQ	-0.47	1.46
100-199	on-t.2.EL	0.04	-0.60
100-199	on-t.2.PE	0.27	-0.95
100-199	on-t.2.YN	-0.23	1.13
100-199	on-t.3.GE	0.35	-0.78
100-199	on-t.3.PD	-0.02	-0.03
100-199	on-t.3.PE	0.48	-0.42
100-199	on-t.3.YN	-1.06	3.13

Appendix B: Displacement values and Robust Z statistics of the 168 continuing items

Count	Item	Displacement	Robust Z
100-199	recr.3.GE	-0.09	0.03
100-199	recr.3.YN	0.29	0.26
100-199	tran.1.EL	-0.03	-0.60
100-199	tran.2.GE	0.13	-1.05
100-199	tran.2.YN	-0.15	1.07
100-199	tran.3.GE	0.21	-1.34
100-199	weat.2.PE	-0.03	0.15
200-299	civi.2.CQ	0.15	0.60
200-299	civi.2.PD	0.42	-0.74
200-299	comm.2.CQ	-0.30	1.19
200-299	cons.1.YN	-0.27	1.22
200-299	cons.2.PE	0.13	-0.72
200-299	educ.1.YN	-0.26	1.23
200-299	educ.3.PD	-0.06	-0.03
200-299	fami.1.EQ	-0.06	0.00
200-299	fami.2.GE	0.10	-0.72
200-299	fami.2.YN	-0.25	1.26
200-299	gett.1.EQ	0.08	0.09
200-299	gett.3.YN	-0.11	0.86
200-299	heal.1.CQ	-0.11	1.13
200-299	heal.4.CQ	-0.13	0.92
200-299	heal.4.PE	0.22	-0.95
200-299	hous.2.eq	-0.18	0.90
200-299	on-t.2.PD	0.31	-0.38
200-299	tran.1.PE	0.11	-0.65
200-299	tran.1.YN	-0.76	1.97
200-299	tran.2.CQ	-0.28	1.10
200-299	tran.2.EL	0.39	-1.36
200-299	tran.3.CQ	-0.17	0.66
200-299	tran.3.YN	0.20	0.83
200-299	weat.2.CQ	-0.36	1.36
300 or more	civi.2.EQ	0.08	0.53
300 or more	comm.2.EQ	0.09	0.24
300 or more	comm.2.PD	0.14	-0.21
300 or more	comm.2.YN	0.24	0.59
300 or more	cons.1.CQ	-0.27	1.16
300 or more	cons.1.EQ	0.03	0.32
300 or more	cons.1.PD	0.11	-0.23
300 or more	cons.2.EL	0.16	-0.59
300 or more	cons.2.EQ	0.04	0.57

Appendix B: Displacement values and Robust Z statistics of the 168 continuing items

Count	Item	Displacement	Robust Z
300 or more	cons.2.PD	0.10	-0.09
300 or more	educ.1.CQ	-0.39	1.25
300 or more	educ.1.EQ	0.17	-0.08
300 or more	educ.1.PD	-0.12	0.14
300 or more	educ.1.PE	0.31	-0.54
300 or more	educ.3.EQ	0.24	-0.26
300 or more	educ.3.PE	0.19	-0.89
300 or more	educ.3.YN	0.33	0.27
300 or more	fami.2.CQ	-0.37	1.29
300 or more	fami.2.EQ	-0.35	1.05
300 or more	fami.2.PD	0.10	-0.21
300 or more	fami.2.PE	0.21	-0.74
300 or more	gett.3.EQ	0.14	-0.02
300 or more	heal.1.EQ	0.06	0.18
300 or more	heal.1.PD	-0.11	0.08
300 or more	heal.1.PE	0.25	-0.41
300 or more	heal.1.YN	-0.25	1.14
300 or more	heal.2.CQ	-0.37	1.28
300 or more	heal.2.EQ	-0.03	0.75
300 or more	heal.2.GE	0.24	-0.71
300 or more	heal.2.PD	-0.06	0.08
300 or more	heal.4.EQ	0.05	0.68
300 or more	heal.4.PD	-0.12	0.11
300 or more	on-t.2.EQ	-0.31	1.17
300 or more	on-t.3.EQ	-0.15	0.39
300 or more	recr.3.CQ	0.02	0.78
300 or more	recr.3.EL	0.19	-0.36
300 or more	recr.3.EQ	0.12	0.06
300 or more	recr.3.PD	-0.01	-0.06
300 or more	recr.3.PE	-0.01	0.33
300 or more	tran.1.CQ	-0.35	1.29
300 or more	tran.1.EQ	0.07	0.20
300 or more	tran.1.PD	-0.03	0.02
300 or more	tran.2.EQ	-0.02	0.44
300 or more	tran.2.PD	0.01	0.00
300 or more	tran.3.EQ	0.06	-0.24
300 or more	tran.3.PD	0.09	-0.32
300 or more	tran.3.PE	0.19	-0.50
300 or more	warmup.2.WA	-0.81	1.90
300 or more	warmup.3.WA	-0.35	1.17

Appendix B: Displacement values and Robust Z statistics of the 168 continuing items

Count	Item	Displacement	Robust Z
300 or more	warmup.4.WA	-0.28	1.04
300 or more	warmup.5.WA	0.00	-0.30
300 or more	warmup.6.WA	-0.63	1.49
300 or more	warmup.7.WA	0.23	-0.62
300 or more	weat.2.EQ	0.06	-0.26
300 or more	weat.2.GE	0.38	-0.47
300 or more	weat.2.PD	0.02	-0.12
300 or more	weat.2.YN	-0.38	1.28

CAL
CENTER FOR APPLIED LINGUISTICS

www.cal.org

Visit www.cal.org/aea/bestplus/2/ to learn more about BEST Plus Version 2.0.