UCLA | CRESST NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

# ENSURING THE COMPARABILITY OF MODIFIED TESTS ADMINISTERED TO SPECIAL POPULATIONS

**Phoebe C. Winter, Mark Hansen, and Michelle McCoy**

JUNE 2019

CRESST REPORT **864**

This research and methodology development were performed to enhance the ELPA21 assessment system and better serve English learners who are blind or have low vision.



To cite from this report, please use the following as your APA reference: Winter, P. C., Hansen, M., & McCoy, M. (2019). *Ensuring the comparability of modified tests administered to special populations* (CRESST Report 864). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

# Table of Contents

# Ensuring the Comparability of Modified Tests Administered to Special Populations[1]

Phoebe C. Winter

Independent Consultant

Mark Hansen and Michelle McCoy

ELPA21 @ UCLA/CRESST

**Abstract:** In order to accurately assess the English language proficiency of special populations of English learners, student assessment programs must maintain the comparability of standard and modified assessment formats, allowing for equivalent inferences to be made across student classifications. However, given the typically small size of special populations of English learners, such as blind and low vision students, traditional calibration and item linking techniques are often incapable of ensuring sufficient levels of comparability. With this in mind, researchers at CRESST set out to test the efficacy of a new item calibration technique: one in which the overall cut scores on the ELPA21 braille form require the same language skills and knowledge as the cut scores on the ELPA21 online form.

To test their approach, the researchers recruited a panel of six educators who specialize in working with special populations of students. Using data from the 2016-2017 braille test administration, the panelists reviewed the similarities and expected differences in difficulty between test form items. They went on to estimate the proportion of students taking the braille form who would meet the target student descriptors and correctly responded to test items. In doing so, the panel was able to successfully recommend cut scores for the braille form while preserving the necessary comparability with cut scores on the online form. This process could be utilized for other assessments in which modified test items are present and cannot be said to have the same parameters as their source items.

[1]This report was originally presented at the annual meeting of the National Council on Measurement in Education, April 2018, New York, NY.

# Introduction

Most student assessment programs provide their tests in more than one format in order to access estimates of achievement or proficiency for all students. For example, a paper-based test may be offered in large print or an online test offered on paper. In most cases, changes in item format do not appreciably alter the item content, and the item calibrations or raw-to-scale lookup tables for the base form are used to provide scale scores for the alternative format test. In other cases, however, changes in test formats may be reasonably expected to alter the difficulty of items. Changes from an interactive, online format to a static, paper format may result in different item parameters, depending on the nature of the revisions. Item substitutions, which may be necessary when changing delivery formats, will certainly affect raw to scale score conversions.

Research has increasingly addressed methods for linking scores from tests delivered in different formats, particularly for paper-based and computer-delivered assessments (e.g., Council of Chief State School Officers, 2017; Dady, Lyons, & DePascale, 2018; Eignor, 2007; Lottridge, Nicewander, & Mitzel, 2010; Steedle, McBride, Johnson, & Keng, 2016). Some attention has been paid to deriving equivalent scores on tests that measure the same constructs in alternative formats and also contain some modified items (e.g., Adams, 2007; Dorans, Pommerich, & Holland, 2007; Sireci & Wells, 2010; DePascale, 2010; Evans & Lyons, 2017).

The techniques proposed by these authors to link scores from different formats rely on large samples of examinees. Often, test forms presented in an alternative format are not administered to enough examinees to allow for statistical score linking techniques. This presents a significant barrier to reporting scores for examinees taking the alternative format that are comparable to scores for examinees taking the general form of the test. When tests are used for high-stakes decisions such as student placement or program evaluation, the ability to make equivalent inferences across formats is critical.

State academic and language proficiency assessments typically use test scores to classify students into ordered groups for the purpose of describing student status on the construct assessed, determining appropriate placement or educational options for students, and characterizing school and district status and progress in educating students. In these cases, inferences about student achievement or proficiency are not made based on test score; they are made based on student classification. Of interest in this situation is providing comparable classifications of students. That is, comparability is defined as the ability to make the same inferences about the category to which a student's performance corresponds.

Mislevy (1992) describes five different types of score linking, from equating to social moderation, and discusses their applicability based on the type of inferences made from the assessments. Linn's (1993) social moderation category of linking is appropriate in the case of an assessment covering the same construct but in a different format, with some different items, a

small number of examinees, and the desired inference made about the examinee's achievement/proficiency level. Social moderation requires judgments about the comparability of student performance across the assessments being considered.

This report describes the tryout of a procedure for linking cut scores on tests with different formats that are designed to measure the same construct and a trial implementation of the process. ELPA21 measures the academic language proficiency of students who are English learners in kindergarten through Grade 12 to determine eligibility for English language development services and monitor English learners' annual progress. Proficiency is measured in each of four domains—reading, listening, writing, and speaking. The test is based on the *English Language Proficiency Standards* (Council of Chief State School Officers, 2014), which were developed by linguistics and education experts at CCSSO, WestEd, the Understanding Language initiative at Stanford University, and state departments of education. English language proficiency tests have been developed for students in specific grades or grade bands: kindergarten, Grade 1, Grade Band 2–3, Grade Band 4–5, Grade Band 6–8, and Grade Band 9–12 (HS).

Because one of the goals of the test is to obtain measures of language proficiency for each domain, it was developed to minimize the potential interference (construct-irrelevant variance) that proficiency or degree of proficiency in one domain may have on the measure of proficiency in another domain. One way of doing this was to incorporate graphics (and to a lesser extent, videos) as prompts and supports in ELPA21 tasks, resulting in the test relying relatively heavily on visual prompts and stimuli.

In order to make the test accessible for students who are blind or have low vision, individuals with expertise in this population reviewed test items and provided suggestions concerning appropriate modifications to items and tasks. Some task types were deemed suitable for administration in braille with no additional changes. Others required further revision, such as eliminating graphics in task prompts or replacing some types of visuals with others more easily rendered in braille; in some cases, the response mode changed. For some task types, no revisions were possible, so tasks measuring the same constructs were developed for the braille forms; these new tasks included some items that use manipulatives.

Implementation of these changes produced braille tests that were judged to assess the same language skills as the online tests taken by other students. However, the process of revising the presentation and response modes for the items meant that item parameters from the online calibration would be of questionable accuracy for the braille forms. Since a very small number of students require braille testing, standard calibration or linking methods could not be applied. Yet it is critical that performance on these alternative forms support the same inferences about student proficiency as the forms administered to other students.

As is the case with ELPA21, state-level academic assessments do not typically have enough students taking the braille version of the test to conduct separate item and test

calibration. In many, if not most, state assessment programs, the development of braille forms uses the existing paper-and-pencil or online item pool as its basis. Items are selected for the braille form to match the test blueprint, with the additional requirement that the items be relatively easy to translate from their base version to a braille presentation format. The braille forms are then scored using the item calibrations from their source items and using the same scaling procedures used on the general, non-accommodated forms.[2]

As described earlier, the design of the braille form of ELPA21 differs from that of most state academic assessment braille forms. Many of the items do not have direct corollaries on the online form. While the design of the braille form might lend itself to a traditional score linking process, using the items translated relatively directly into braille as common items, the number of students taking the braille form—a total of 23 examinees across all grades and grade bands during school year 2016–2017—precluded such a procedure.

## Method

The goal of this linking approach is to have overall cut scores on the braille forms that require the demonstration of the same level of knowledge and skills as the cut scores on the online forms. In essence, the task may be viewed as one of translating the previously established cut scores for the online form onto the raw scale (i.e., summed score) of the braille forms. This study describes a tryout of the procedure with the purpose of troubleshooting specific information and processes used while providing a proof of concept for the actual implementation of the approach in a later study.

The procedure that was developed and tried out relies on expert judgment, informed by data, in order to set linked cut scores on Levels 3 and 4, out of 5 possible score levels, on each of the grade-band specific braille-version domain tests. These cut scores can be used to derive an overall cut score for English language proficiency using the same rules that are in effect for the online test: *proficient* = scores of 4 or higher on all domains; *progressing* = at least one score 3 or above and at least one score below 4; *emerging* = scores of less than 3 on all domains.

To provide context, a description of the standard setting process for the online form follows.

## Standard Setting on the Online Form

Cut scores for the ELPA21 online forms were set after the first operational administration of the test in 2015–2016. ELPA21 has tests for each of the four domains (reading, listening, writing, and speaking) for kindergarten, Grade 1, Grade Band 2–3, Grade Band 4–5, Grade Band

[2]Based on personal communication from several state and consortium assessment staff members: Bob Lee, MA; Steve Slater, OR; Shaun Bates, MO; Joseph Saunders, SC; Kara Todd, WA; and Matthew Schulz, Smarter Balanced (January, 2018).

6–8, and Grade Band 9–12. Cut scores have been established at each of five levels (1 = *beginning*, 2 = *early intermediate*, 3 = *intermediate*, 4 = *early advanced*, and 5 = *advanced*) for each grade, K through 8, and for the 9–12 (HS) grade band.

The Bookmark procedure (Lewis, Mitzel, Mercado, & Schulz, 2012) was used to recommend cut scores for Levels 3 and 4 for each domain in each grade or one grade of the grade band test—K, 1, 3, 5, 7, and HS—these scores were adjusted as needed to achieve proper ordering based on the relationship between item parameters for common items in adjacent grades. The final Level 3 and 4 cut scores for Grades K, 1, 3, 5, and 7 were used to set cut scores for Levels 3 and 4 in Grades 2, 4, 6, and 8 at the midpoint between the cut scores at the grades below and above. Cut scores for Levels 2 and 5 were set based on creating approximately equal-sized groups in Levels 1 and 2 and approximately equal-sized groups in Levels 4 and 5. (See Pacific Metrics & UCLA CRESST, 2016, *ELPA21 Standard Setting Technical Report* for more detail [available upon request].)

Overall English language proficiency is determined based on the levels that students reach on the domain forms. If the student earns domain scores in level 4 or higher on all domains, the student is classified as *proficient*. Earning at least one domain score below the cut score for level 4 and at least one score in level 3 or higher classifies a student as *progressing*. If the student's domain scores are all in level 2 or below, the student is classified as *emerging*.

# Participants

The linking approach was tried out with a group of educators with expertise in educating students with disabilities, students who are English learners (ELs), and students who are blind or have low vision. Six panelists participated in the trial run, which was facilitated by one of the developers of the procedure. ELPA21 staff were available to answer questions about the assessment and materials. The panel composition was as follows:

- Gender: four female, two male
- Race/ethnicity: four White, non-Hispanic; two Asian/Pacific Islander
- Primary area of expertise: one educating students with disabilities in general; two educating students who are blind; two educating ELs in general; one educating students who are blind, some of whom are ELs.

# Materials

## Braille Test Forms

The ELPA21 braille forms are designed for students who are blind or have low vision and cannot take the test with accommodations provided on either the online or paper form; students taking the braille form are learning or already read braille. The ELPA21 braille forms closely conform to the online blueprint in terms of task type, standards coverage, and

standards emphasis so that claims and inferences made about student English language proficiency on these forms can be comparable to those made on the online form. Each braille version test consists of tasks and items that correspond to those offered on the online form in terms of both academic context and knowledge and skills assessed. This approach allows ELPA21 to avoid extrapolating from an incomplete measure of the standards in generating scores for these students, an approach that is often necessitated when only direct braille and audio conversions are used to represent the targeted English language proficiency constructs for this population.

As part of item development, each item in the ELPA21 online item pool was evaluated by content and accessibility experts to determine whether the item was accessible to students who are blind, either as is or by using a braille version. The pool of accessible items was then reviewed to identify gaps in standards coverage so that items could be revised or additional items could be developed to fill in the gaps. Existing items and tasks were revised or new items and tasks were developed, using online items and tasks as models. This created an initial pool of items accessible to students who are blind. For example, a reading item with a pictograph was delivered in braille and revised to include a table containing the same information as the pictograph. A substitute task for a listening item with picture options was developed to use manipulatives instead of the picture options. The resulting pool of items eligible for the braille form was reviewed by external experts to ensure that the items measured the targeted constructs appropriately for the given student population.

In the braille form, as in the online form, care was taken to ensure that, as much as possible, the constructs of the four domains did not overlap. For example, if in the online writing test the prompt is available both orally and in text to minimize the effects of reading skills in understanding the task, then the prompt in the braille form is delivered both orally and in braille.

In preparation for the trial run, the relationships between the braille form items and their corollaries on the online form were codified based on their similarity. This scheme was used in the linking procedure and is also useful for understanding the nature of the braille forms:

1. The braille item version is identical to the online item, other than necessary differences in presentation format; changes are superfluous to apprehending the meaning of the item.
2. Similar to Level 1, but the response format is modified, or a visual that orients students to the text has been deleted. For example, pictures that accompany a read-aloud story and do not contribute to comprehending the story are dropped.
3. While the braille form items are based on the online item, revisions may cause the depth of the concept measured to be different from the online versions. For example, a pie chart in the online version is changed to a table.

4. These braille form items are designed to measure a concept from the ELPA21 standards and Achievement Level Descriptors (ALDs), but the items have different content. For example, an online item may ask students to compare two pictures, while the modified item measuring the same standards asks the student to compare two events.

The braille tests are administered individually by a trained test administrator familiar to the student, typically the student's teacher or vision specialist. Students respond to selected response items by telling the administrator which option(s) they select and to constructed response items on the writing assessment using a braillewriter or slate and stylus.

The braille test forms used in the trial run were administered in school year 2016–2017. Panelists reviewed all four domain forms in Grade 1 and Grade Band 4–5 so that a variety of item types and levels would be included in the trial. Panelists also reviewed the Grade Band 6–8 reading form, since it contained longer passages to evaluate the time needed to review upper grade level reading forms and ensure that the planned procedures worked in the same way on a higher grade band form. The other three domains at the upper grade levels are similar enough in task type composition to Grade Band 4–5 that they were not reviewed as part of the trial run.

For the braille items with an original online item source, panelists were provided with a text-based description of the braille item and a description of how it had been adapted from its online source. For braille items without an online source, panelists were provided with the text-based description of the item. The actual braille versions of the forms and associated manipulatives were available for reference.

## Ancillary Materials

Prior to the in-person meeting, panelists were provided with the following materials:

- achievement level indicators (ALIs), which describe the skills eligible to be tested, for Grade 1 and Grade Band 4–5, all domains and for Grade Band 6–8, reading only
- a chart characterizing the relationships between items on the braille and online forms on a 1 to 4 scale (see Table 1); while panelists were given this rubric as a guide to help them estimate braille item difficulties at Proficiency Levels 3 and 4, they were encouraged to add their own interpretations of item functioning in the two populations when making their judgments
- target student descriptors (TSDs) for Levels 3 and 4 used during standard setting, which describe the skills of students in each proficiency level, for Grades 1 and 5, all domains, and Grade 7, reading only

Table 1

*Relationships Between Braille and Online Items (Levels of Similarity and Expected Differences in Difficulty)*

| Group | Definition (relationship with source) | Item difficulty |
|---|---|---|
| 1 | Identical to the online item, other than necessary format of presentation; changes are superfluous to apprehending the meaning of the item. That is, we do not expect significant trait/format interactions. | Item difficulties should be close to the online item difficulty. |
| 2 | Similar to Level 1, but the response format is modified or a visual that orients students to the text has been deleted. For example, pictures that accompany a read-aloud story are dropped. | Item difficulties should be closely related to online, but may not be as close as a #1 item's difficulty. |
| 3 | While these items are based on the online item, modifications cause the depth of concept measured to be different from the online versions. For example, a graph in the online version is changed to a table. | The relationship to the difficulty of the online source item will vary, depending on the specific changes made. |
| 4 | These items are designed to measure a concept from the ELPA21 standards and ALDs, but the items have different content. For example, an online item may ask students to compare two pictures, while the modified item measuring the same standards asks the student to compare two events. | There is no consistent relationship expected between item difficulties. |

During the in-person meeting, panelists were provided with the following materials:

- hard copy of the ALIs and TSDs
- hard copy of the rating scale for relationships between items
- a sample review form in Excel
- hard copy of sample items from the Grade Band 4–5 test, with scoring rubrics when applicable

Also available for examination during the in-person trial run were braille student test booklets, the Directions for Administration (DFA) documents for all grade band tests, and the manipulative kits for each grade band.

## Performance Data

For each braille item with an original online source, panelists were provided with the expected percent correct or the percent assigned each score point for polytomous items for examinees at the previously established cut points of Levels 3 and 4 on the online test. An illustration is shown in Figure 1.

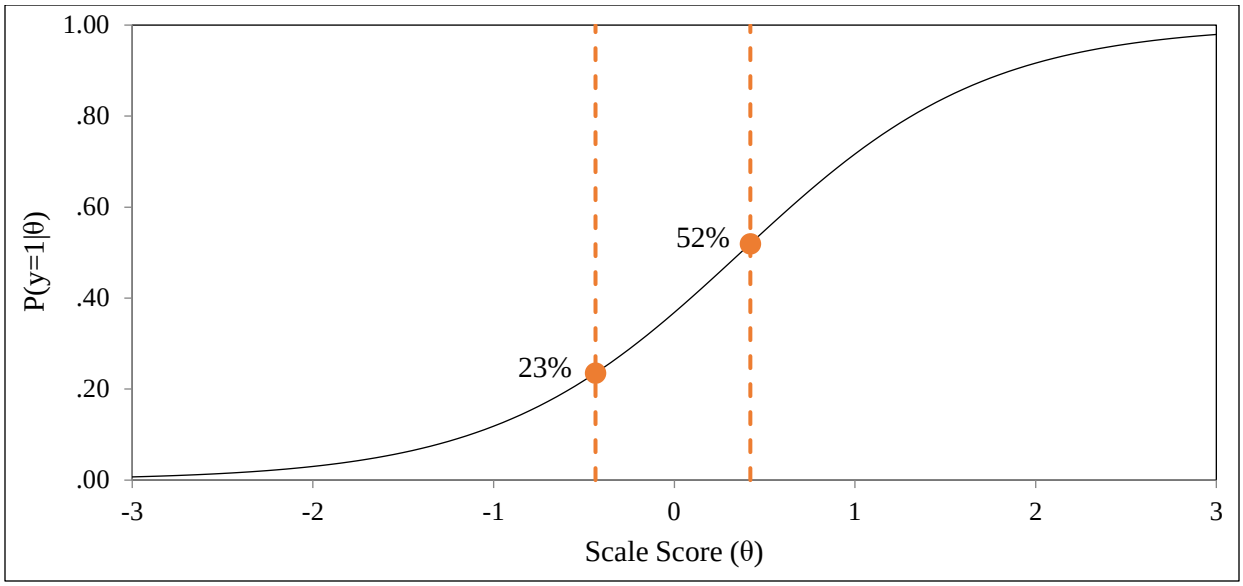


*Figure 1.* Traceline and expected percent correct at cut scores for a Grade 1 reading item at Levels 3 and 4. Note: The traceline shows the percent correct as a function of the scale score. The vertical lines indicate the cut scores at which the percent correct was evaluated.

Table 2 describes the forms and associated data used in the trial run. Most braille items with a relationship of 1, 2, or 3 with their online counterpoints had data available that showed the expected performance of students scoring just at the online cut score for Levels 3 or 4.

Table 2

*Trial Run Braille Test and Item Characteristics*

| | | | | | Relationship | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Grade/domain | Points | Items[a] | Dichotomous | Polytomous | 1 | 2 | 3 | 4 | Online data |
| Grade 1 | | | | | | | | | |
| Reading | 23 | 23 | 23 | 0 | 12 | 3 | 4 | 4 | 19 (83%) |
| Listening | 21 | 21 | 21 | 0 | 8 | 3 | 0 | 10 | 11 (52%) |
| Writing | 21 | 8 | 0 | 8 | 0 | 4 | 3 | 1 | 6 (75%) |
| Speaking | 16 | 5 | 0 | 5 | 1 | 2 | 0 | 2 | 3 (60%) |
| Grade Band 4–5 | | | | | | | | | |
| Reading | 24 | 23 | 22 | 1 | 16 | 0 | 3 | 4 | 19 (83%) |
| Listening | 30 | 27 | 22 | 5 | 7 | 9 | 0 | 11 | 15 (56%) |
| Writing | 27 | 14 | 9 | 5 | 10 | 0 | 3 | 1 | 6 (43%) |
| Speaking | 31 | 10 | 0 | 10 | 5 | 0 | 1 | 4 | 4 (40%) |
| Grade Band 6–8 | | | | | | | | | |
| Reading | 31 | 28 | 26 | 2 | 16 | 7 | 5 | 0 | 23 (82%) |

[a]Item = a unit of the test that receives a single score; in some cases, an item may require responses to multiple prompts.

## Procedures

Before the in-person meeting, panelists participated in a webinar that included

- an overview of the purposes and goals of the trial run, including a brief overview of ELPA21 standard setting
- a description of what ELPA21 measures and how the braille form was developed
- an explanation of the relationships between braille form items and online form items (see Table 1), with examples
- a description of the braille forms and how they are administered
- an overview of target student descriptors used in standard setting

Panelists met in person to link the cut scores on the online forms to performance on the braille forms. They first met as a single group for additional orientation and training in reviewing and rating the braille items. Topics covered were:

- revisiting the purpose of the task and description of forms,
- revisiting the ALIs and TSDs,

- revisiting the relationships between online and braille items and the 1–4 item designation,
- a discussion of questions arising after the webinar,
- a more detailed discussion of how standards were set and the meaning of cut scores,
- an overview of the review process and panelist roles,
- review and discussion of sample items, including polytomous, rubric-scored items, and
- application of the process using the sample items.

Panelists reconvened into two groups of three, one group reviewing Grade 1 domain forms to determine the cut scores at Levels 3 and 4 for Grade 1 students and the other group reviewing Grade Band 4–5 forms to determine the cut scores for Grade 5 students. Each domain form was reviewed/rated and then discussed, beginning with reading, followed by listening, writing, and speaking.

## Round 1

For each domain, panelists first independently reviewed and rated the items. Panelists were provided with a written description of the items on the braille form. Each item was accompanied by an explanation of how it differed from the source item on the online form (e.g., pictures embedded in a story online were deleted from the braille version), if a source item was available. One copy of each actual braille form was available for panelists to consult if they had questions about the format used in the administration.

Panelists were provided with a review form in Excel that contained the following information:

- the relationship between the braille and source item on the 1 to 4 scale (1 indicating close relationship, 4 indicating no corresponding source item as shown in Table 1)
- for most items with a relationship of 1, 2, or 3[3]
  - the proportion of online students who just met the score for Level 4 expected to get the item correct (or receive each score point, for polytomous items)
  - the proportion of online students who just met the score for Level 3 expected to get the item correct (or receive each score point, for polytomous items)
- a space for panelists to enter their estimate of the proportion of blind/braille users who just met the criteria for Level 4 that would answer the braille item correctly (or receive each score point, for polytomous items)
- a space for panelists to enter their estimate of the proportion of students who are blind and just met the criteria for Level 3 who would get the braille item correct (or receive each score point, for polytomous items)

[3]In some cases, the online proportion was not available due to scoring conventions.

- a space for panelists to enter comments about the items

Panelists were tasked to review each pair of items and determine the percentage of students taking the braille form who had the knowledge and skills to just meet the description in the TSD for Level 4 that would get the item correct (or, in the case of polytomous items, earn each score point). Panelists repeated this task for students who had the knowledge and skills in the TSD for Level 3.

Panelists referred to the performance of online students, the similarity between the braille form item and its online source (when available), and the TSDs when making their estimates of the performance of students on the braille form. When there was no online source, panelists used their understanding of the performance of students who are blind and considered their estimates of performance on other tasks to make estimates of student performance on the braille form item.

### Round 2

At each table, a panelist who was familiar with ELPA21 design and development served as a table leader and facilitated the discussion during Round 2. The sum of a panelist's Round 1 estimates (with appropriate weighting for polytomous item scores) was calculated to determine the panelist's initial raw cut score recommendations for Level 4 and for Level 3. Panelists were shown the range of recommended scores and discussed the potential scores, with the goal of reaching consensus on a final recommendation.

In case panelists disagreed widely on their recommended cut scores, procedures were planned to have panelists discuss items on which they were most discrepant in their ratings. These procedures were designed to surface how panelists were defining student knowledge and skills and to discuss their thoughts on the effects of item changes on the constructs for students taking the braille form. After the discussion, panelists could revise any item ratings they desired on their spreadsheets, which enabled them to see the potential effects that their revised ratings would have on their overall cut score as they made the changes. These revised ratings would be discussed to reach consensus on a final recommendation. During the trial, this second item rating was not needed, since panelists easily came to consensus. Although there were no large disagreements, panelists did discuss the most discrepant item in Grade Band 6–8 so that the discussion procedure could be tried out.

Round 1 and Round 2 were repeated for each domain.

## Results

Because this was a trial run and panelists were able to reach consensus easily, no formal record of the final recommended cut scores was kept. After discussion, panelists often recommended the rounded median cut score. When panelists felt there was not enough separation between the median cut scores, they discussed which cut score or scores should be

changed and why. In some cases, panelists lowered the Level 4 cut score if it was too close to the maximum number of obtainable points. Table 3, Table 4, and Table 5 show panelists' initial Round 1 recommendations. The appendix contains item-by-item results for the Grade 1 reading (all dichotomous) and Grade 1 speaking (all polytomous) reviews. The degree of agreement was similar in other domains and grade bands.

Table 3

*Grade 1 Recommended Cut Scores*

| | Domain (max points) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Reading (23) | | Listening (21) | | Writing (21) | | Speaking (16) | |
| Panelist | L3 | L4 | L3 | L4 | L3 | L4 | L3 | L4 |
| Panelist A | **13.42** | 16.81 | 16.64 | 18.90 | 17.05 | **18.42** | **13.45** | **13.97** |
| Panelist B | 13.17 | **16.84** | 14.18 | 18.22 | 16.28 | 18.20 | 13.43 | 13.89 |
| Panelist C | 13.83 | 17.09 | **15.53** | **18.54** | **16.95** | 18.97 | 13.61 | 14.34 |
| Range | .41 | .28 | 1.09 | .35 | .77 | .77 | .18 | .45 |

*Note.* Median is bolded.

Table 4

*Grade 5 Recommended Cut Scores*

| | Domain (max points) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Reading (24) | | Listening (21) | | Writing (21) | | Speaking (16) | |
| Panelist | L3 | L4 | L3 | L4 | L3 | L4 | L3 | L4 |
| Panelist D | **12.22** | 17.52 | 18.72 | 23.56 | 19.00 | 23.66 | 23.37 | 27.63 |
| Panelist E | 11.03 | **17.83** | **18.90** | **24.19** | 18.23 | **24.18** | **22.97** | 25.70 |
| Panelist F | 12.79 | 18.22 | 21.27 | 24.36 | **18.81** | 24.19 | 22.92 | **25.88** |
| Range | 1.76 | .70 | 2.55 | .80 | .77 | .53 | .45 | 1.93 |

Table 5

*Grade 7 Recommended Cut Scores*

| | Domain (max points) | |
|---|---|---|
| | Reading (24) | |
| Panelist | L3 | L4 |
| Panelist A | **19.27** | 24.98 |
| Panelist B | 18.00 | 23.70 |
| Panelist C | **19.50** | **26.01** |
| Panelist D | 20.21 | 26.58 |
| Panelist E | 18.19 | 26.93 |
| Panelist F | 20.03 | **25.72** |
| Range | 2.03 | 3.23 |

*Note.* The two middle scores are bolded.

## Discussion

The purpose of the trial run was to evaluate the proposed procedure to see if (a) it was feasible to use this method to set linked cut scores on the braille form and (b) if so, what refinements to the method were needed. Overall, panelists were able to make the required judgments and felt that the task was reasonable, and they were able to reach consensus on a proposed cut score at each level. Refinements recommended to the process by participants and the research team ranged from suggestions for logistics (e.g., type of spreadsheet used) to suggestions for procedures (e.g., providing a demo of braille administration). The method itself—determining the percentage of students taking the braille form who would be expected to get an item correct (or earn a specific number of points), using online student performance as a reference—was deemed feasible, and the recommended changes to the procedures and materials will enhance but not change their essential nature.

The major substantive change to the method based on the trial run results is to train the panelists in the 4-point item similarity scale and to emphasize that their expert judgment of item similarity was critical by not labeling items with the similarity rubric. A number of modifications to logistics, materials, and procedures are planned based on participation evaluation results. These include changes such as providing additional detail about standard setting on the online form and providing panelists with the rounded median score after Round 1.

This process could be used with other assessment variations that incorporate modified items that cannot be assumed to have the same item parameters as their source items. For example, paper-based versions of online tests necessarily include modified formats for drag-and-drop items. Often, these items use a grid to include the same number and nature of choices for students to respond to. However, these modifications may change the difficulty of the items. Similarly, tests that are transadapted to other languages may also incorporate changes to items that cannot be assumed to leave item difficulty the same. More generally, the procedure can be used in high-stakes situations that employ human judgment in decision making. The use of available statistics can help ground the panelists as they make their decisions.

## Limitations

In this pilot, there were no methods for evaluating the quality of the results beyond soliciting feedback from the panelists. Evaluation methods used in standard setting, for example, intra- and inter-judge consistency, do not apply, since panelists are presented with a base percent correct, constraining their responses by design. Qualitative evaluations based on panelists' comfort with procedures can be used and were part of this study. However, there is a need to devise methods to evaluate the process. Replication using multiple panels was not used in the pilot due to lack of sufficient experts to serve on panels. If possible, future applications of the method should incorporate multiple panels.

Applying this procedure to alternative test formats with larger numbers of examinees, for example, computer/device-based items delivered on paper, would allow us to evaluate results by investigating actual examinee responses on the paper-based test and comparing them to panelist judgments.

# References


Adams, R. (2007). Cross-moderation methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 212–245). London, England: Qualifications and Curriculum Authority.

Council of Chief State School Officers. (2014). *English language proficiency standards.* Washington, DC: Author.

Council of Chief State School Officers. (2017). *Score comparability across computerized assessment delivery devices.* Washington, DC: Author.

Dady, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. *Applied Measurement in Education, 31,* 30–50.

DePascale, C. (2010). Modified tests for modified achievement standards: Examining the comparability of scores to the general test. In P/C Winter (Ed.), *Evaluating the comparability of scores from the educational achievement test variations*. Washington, DC: Council of Chief State School Officers.

DePascale, C. (2010a). Evaluating linguistic modifications: An examination of the comparability of a plain English mathematics assessment. In P.C. Winter (Ed.), Evaluating the Comparability of scores from educational achievement test variations. Washington, DC: Council of Chief State School Officers.

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). The descent of linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 355–359). New York, NY: Springer Science + Business Media.

Eignor, D. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), Lin*king and aligning scores and scales* (pp. 135–159). New York, NY: Springer Science + Business Media.

Evans, C. M., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. *Educational Measurement: Issues and Practices, 36,* 24–34.

Lewis, D.M., Mitzel, H.C., Mercado, R. L., Schulz, E.M. (2012). The bookmark standard setting procedure. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations.* New York, NY: Routledge.

Linn, R. 1993. “Linking Results of Distinct Assessments.” *Applied Measurement in Education* 6(1).

Lottridge, S., Nicewander, W. A., & Mitzel, H. (2010). Summary of the online comparability studies for one state’s end-of-course program. In P. Winter (Ed.), *Evaluating the*

*comparability of scores from achievement test variations* (pp. 13–32). Washington, DC: Council of Chief State School Officers.

Mislevy, R. L. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS.

Pacific Metrics, UCLA CRESST. (2016). *ELPA21 Standard Setting Technical Report*. Los Angeles, CA: University of California.

Sireci, S. G., & Wells, C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33–68). Washington, DC: Council of Chief State School Officers.

Steedle, J., McBride, M., Johnson, M., & Keng, L. (2016). *Spring 2015 digital devices comparability research study.* Iowa City, IA: Pearson.

# Appendix: Sample Item-Level Results

Grade 1 Reading

| | | | Level 4 | | | | Level 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Position | Relationship | Max Score Pts | Online | Panelist A Estimate | Panelist B Estimate | Panelist C Estimate | Online | Panelist A Estimate | Panelist B Estimate | Panelist C Estimate |
| 1 | 3 | 1 | 100% | 100% | 100% | 100% | 97% | 97% | 97% | 97% |
| 2 | 3 | 1 | 100% | 100% | 100% | 100% | 95% | 95% | 95% | 95% |
| 3 | 3 | 1 | 99% | 100% | 99% | 99% | 92% | 90% | 92% | 92% |
| 4 | 3 | 1 | 99% | 97% | 99% | 99% | 94% | 90% | 94% | 94% |
| 5 | 4 | 1 | NA | 95% | 90% | 90% | NA | 85% | 85% | 85% |
| 6 | 4 | 1 | NA | 95% | 90% | 95% | NA | 90% | 85% | 90% |
| 7 | 4 | 1 | NA | 100% | 95% | 100% | NA | 90% | 90% | 90% |
| 8 | 4 | 1 | NA | 90% | 95% | 90% | NA | 90% | 95% | 90% |
| 9 | 1 | 1 | 91% | 85% | 85% | 90% | 68% | 60% | 60% | 68% |
| 10 | 1 | 1 | 69% | 69% | 60% | 69% | 49% | 40% | 40% | 40% |
| 11 | 1 | 1 | 85% | 75% | 75% | 75% | 64% | 55% | 55% | 55% |
| 12 | 1 | 1 | 71% | 65% | 71% | 60% | 43% | 35% | 43% | 43% |
| 13 | 2 | 1 | 61% | 50% | 50% | 50% | 29% | 20% | 20% | 20% |
| 14 | 2 | 1 | 63% | 60% | 50% | 63% | 48% | 45% | 20% | 48% |
| 15 | 2 | 1 | 66% | 60% | 55% | 66% | 59% | 50% | 45% | 50% |
| 16 | 1 | 1 | 74% | 67% | 74% | 70% | 52% | 47% | 52% | 50% |
| 17 | 1 | 1 | 53% | 50% | 53% | 53% | 40% | 40% | 35% | 40% |
| 18 | 1 | 1 | 58% | 55% | 58% | 58% | 47% | 45% | 47% | 47% |
| 19 | 1 | 1 | 63% | 55% | 60% | 60% | 33% | 30% | 25% | 30% |
| 20 | 1 | 1 | 52% | 50% | 50% | 52% | 23% | 20% | 10% | 23% |
| 21 | 1 | 1 | 43% | 43% | 43% | 43% | 37% | 37% | 37% | 37% |
| 22 | 1 | 1 | 57% | 55% | 57% | 57% | 44% | 42% | 40% | 44% |
| 23 | 1 | 1 | 69% | 65% | 75% | 70% | 51% | 49% | 55% | 55% |
| Score | | | | 16.81 | 16.84 | 17.09 | | 13.42 | 13.17 | 13.83 |

Grade 1: Speaking

| | | | Level 4 | | | | Level 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Position | Relationship | Score Point | Online | Panelist A Estimate | Panelist B Estimate | Panelist C Estimate | Online | Panelist A Estimate | Panelist B Estimate | Panelist C Estimate |
| 1 | 4 | 0/2 | NA | 0% | 1% | 0% | NA | 1% | 1% | 5% |
| 2 | 4 | 1/2 | NA | 5% | 9% | 5% | NA | 4% | 14% | 15% |
| 3 | 4 | 2/2 | NA | 95% | 90% | 95% | NA | 95% | 85% | 80% |
| 4 | 4 | | | . | | | | . | . | . |
| 5 | 4 | | | | | | | | | |
| Average | | | | **1.95** | **1.89** | **1.95** | | **1.94** | **1.84** | **1.75** |
| 6 | 1 | 0/3 | 9% | 10% | 9% | 5% | 11% | 12% | 11% | 5% |
| 7 | 1 | 1/3 | 12% | 20% | 12% | 10% | 13% | 16% | 13% | 15% |
| 8 | 1 | 2/3 | 36% | 30% | 36% | 40% | 37% | 35% | 37% | 40% |
| | | 3/3 | 43% | 40% | 43% | 45% | 39% | 37% | 39% | 40% |
| Average | | | 2.12 | **2.00** | **2.13** | **2.25** | **2.03** | **1.97** | **2.04** | **2.15** |
| 9 | 2 | 0/4 | 1% | 0% | 2% | 1% | 1% | 1% | 2% | 2% |
| 10 | 2 | 1/4 | 1% | 1% | 2% | 1% | 1% | 1% | 3% | 3% |
| | | 2/4 | 7% | 2% | 10% | 10% | 9% | 11% | 15% | 10% |
| | | 3/4 | 21% | 23% | 21% | 18% | 25% | 20% | 25% | 25% |
| | | 4/4 | 71% | 74% | 65% | 70% | 64% | 67% | 55% | 60% |
| Average | | | 3.60 | **3.70** | **3.45** | **3.55** | **3.49** | **3.51** | **3.28** | **3.38** |
| 11 | 2 | 0/4 | 0% | 0% | 2% | 0% | 1% | 1% | 2% | 0% |
| 12 | 2 | 1/4 | 0% | 0% | 3% | 2% | 1% | 3% | 3% | 1% |
| | | 2/4 | 3% | 3% | 5% | 3% | 4% | 5% | 5% | 4% |
| | | 3/4 | 10% | 12% | 10% | 10% | 13% | 12% | 15% | 15% |
| | | 4/4 | 87% | 85% | 80% | 85% | 82% | 80% | 75% | 80% |
| Average | | | 3.82 | **3.82** | **3.63** | **3.78** | **3.75** | **3.69** | **3.58** | **3.74** |
| 13 | 4 | 0/3 | NA | 5% | 1% | 0% | NA | 8% | 1% | 2% |
| 14 | 4 | 1/3 | NA | 10% | 4% | 1% | NA | 15% | 4% | 5% |
| 15 | 4 | 2/3 | NA | 15% | 10% | 17% | NA | 12% | 20% | 25% |
| 16 | 4 | 3/3 | NA | 70% | 85% | 82% | NA | 65% | 75% | 68% |
| 17 | 4 | | | | | | | | | |
| Average | | | | **2.50** | **2.79** | **2.81** | | **2.34** | **2.69** | **2.59** |
| Score | | | | 13.97 | 13.89 | 14.34 | | 13.45 | 13.43 | 13.61 |

UCLA

CRESST

NATIONAL CENTER FOR RESEARCH ON EVALUATION, STANDARDS, AND STUDENT TESTING

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)

Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522

(310) 206-1532
www.cresst.org