Establishing Comparability between Computer-Based and Paper-Based Formats of the ACCESS for ELLs® Online Summative Assessment

Prepared by:
Shu Jing Yen and James Marcus
November 5, 2014

Issue

As ACCESS for ELLs® moves online as a computer-based (CB) test in academic year 2015-16 (hereafter "ACCESS Online"), there will still be students who take a paper-based (PB) format of the assessment in all four domains. This document presents our plan to address how we will ensure that performances and scores on the two formats will be comparable, particularly during the first operational year (academic year 2015-16).

Background

For operational ACCESS Online, a PB format will be offered as an accommodation for English language learners (ELLs) who are unable to take the CB format. Although a PB format will be used alongside the CB format, the same claims will be made about ELL performances on both vis-à-vis their current level along the continuum of the development of academic English language proficiency. Thus, it is necessary to provide theoretical rationales and empirical evidence that student performances on the two test formats produce comparable results that may be interpreted in the same way—that is, in terms of the proficiency levels defined by the WIDA ELD Standards.

Issues in Score Comparability in the First Operational Year

Because levels of specificity are critical for this discussion, the details here apply only for the ***first operational year*** of ACCESS Online. However, the principles in this discussion will also apply to subsequent years.

In the first operational year, for the Listening and Reading domains, the PB test will consist of items from ACCESS Series 302, which was operational in academic year 2013-14. As pointed out in other read-ahead documents, the ACCESS Online Listening and Reading domains assess the same construct as the ACCESS Listening and Reading domains. The Listening and Reading items on ACCESS and ACCESS Online have the same item specifications and two formats follow essentially the same test blueprint. So while the actual *items* will be different in the two formats, we believe that the *interpretation* of ELL performances on each will be directly comparable and pose no major issues. Comparability of performances between the operational PB ACCESS format and the CB ACCESS Online format will be investigated, as described in the read-aheads on maintaining the score scale for Listening and Reading.

In the first operational year, for the Speaking domain, the PB format will consist of the ACCESS Online CB summative assessment tasks, adapted to a paper format with a media-delivered audio component. Since the two formats consist of the exact same tasks, ELL performances on the Speaking domain between the two formats should be comparable, though investigations about possible modality and scoring effects will be investigated (described later in this read-ahead).

In the first operational year, for the Writing domain in Grades 4-12, the PB format will consist of the ACCESS Online CB summative assessment tasks, which are retired PB tasks from operational ACCESS, in their paper format. [Reminder: for Grades 1-3, only the PB format for the Writing domain will be available for all test takers, so issues of comparability do not exist in those grades.] Since the two formats consist of the exact same tasks, and these tasks have been used operationally on ACCESS, ELL performances on the Writing domain between the two formats should be comparable.

Method

Listening and Reading

For the Listening and Reading domains, we propose using a within-subject design with no counterbalancing to examine the comparability between performances on the CB and the PB formats. The ACCESS Online Listening and Reading Field Tests provide data needed for the comparability studies since ELLs participating in the field test take both CB ACCESS Online Field Test items and PB operational ACCESS items. Since the major difference between the CB ACCESS Online Field Test and the operational PB ACCESS test is the mode of administration, differences in performances between the two tests can be attributed to mode effect. While the actual items on the ACCESS form used during the ACCESS Online Field Test and the one that will be used as the PB accommodation in the first operational year are different, studies conducted on one form can be generalized to other forms because all retired operational ACCESS forms have the same test specifications and blueprints and have been equated to each other in the operational program.

CAL will use guidance provided by Lottridge et al.'s (2010) review of literature on comparability studies to conduct a series of analyses to evaluate the comparability between the CB and PB formats of the summative test in the domains of Listening and Reading. The planned analyses are summarized briefly in Table 1. Taken together, the evidence summarized in the table should present sufficient evidence that both formats measure the same construct and produce scores that can be considered equivalent.

Table 1
Evidence and analyses for establishing comparability between the PB and CB formats of the summative tests for the Listening and Reading domains

Evidence	Planned Analyses
The test content and content specifications must be the same	Demonstration of same test characteristics, scoring rules, item characteristics, item features, etc.
The scores should have the same factor structure	Factor Analysis
The scores should have the same measurement precision	Comparison of Reliability and Test Information Functions for PB and CB formats
The score distributions should differ only in difficulty, and hence, be equitable	Comparisons of Test Characteristics Curves, mean scores, and score distribution for PB and CB formats

Speaking

As mentioned above, in the first operational year the CB and PB formats of the summative assessment in the Speaking domain will consist of the exact same assessment tasks, the tasks developed for ACCESS Online being adapted to a paper-based format. Unlike in the Listening and Reading domains, however, the ACCESS Online Speaking tasks were not field tested using the PB format. Thus, it is not possible to investigate a mode effect using the ACCESS Online Field Test data. Thus, we propose a special data collection effort and analyses to address comparability issues between the two formats.

Additionally, while students' responses to ACCESS Online Speaking tasks will be centrally scored by trained raters in the operational testing program, responses to the paper-based Speaking summative assessment will be scored locally during the test administration by trained test administrators. Thus, a mode effect for the PB version may be confounded with potential rater effects (local versus central scoring) in the operational program. In order to separate these two effects for investigation, CAL proposes two separate research studies to examine each effect separately. The first study is a mode effect study that investigates potential differences in student performances between the PB and CB formats as a special study in the summer or fall of 2015. The second study is a rater effect study that examines potential differences between local and central scoring for a sample of students who will be taking the PB format in the first operational year of ACCESS Online. Details of each study are described below.

Mode Effect Study

The purpose of the mode effect study is to examine whether students perform differently when taking the ACCESS Online Speaking tasks in the PB and the CB format. This is a small-scale study that aims at providing preliminary information about the comparability of performances on the two formats. Thus the study results may also inform necessary revisions to the administration procedures for future operational administration the PB format of ACCESS Online. A counterbalanced within-subject design will be used. One group of 30 students will respond to assessment tasks for the Speaking domain in a PB format followed by a CB format, while another group of 30 students will respond to the same set of tasks from the Speaking domain in opposite order. Counterbalancing is used to moderate any effects that might arise from test order, such as fatigue, practice, or motivation. It is envisioned that some tasks will be the same in both formats, while other tasks will be parallel (i.e., covering a different topic but written to the same specifications).

For practicality, it is proposed that the study focus on the 1, 4-5, and 9-12 grade-level clusters (with 30 students from each cluster for a total of 90 students). Because students from lower elementary (1-3) groups share similar developmental characteristics, results from Grade 1 students should apply to Grades 2-3 students. Similarly, students from Grades 4-5 and 9-12 are quite different, but results from both should apply to students in Grades 6-8 students.

In the proposed study, student responses to both the PB and CB formats of the Speaking tasks will be digitally recorded and scored by the same group of CAL-trained raters, and each student's response will be double scored. A total of four raters will be used for the study. The rating design will ensure that each rater scores performances elicited by both formats (though they will be blind to which format was used to elicit the response). In any case where the same task appears in both formats, the rating design will ensure that no rater will score the same student on that task under both conditions of elicitations.

Students' raw scores between the two elicitation formats will be compared. Observed differences in student performances across the two formats for both same and parallel tasks will be examined to provide information on the extent to which a mode effect appears to occur.

Rating Condition Effect Study

The purpose of the rating condition effect study is to examine whether there is a significant difference between ratings assigned locally (i.e., live) by test administrators (trained through an online rater training program) to students administered the PB format and ratings assigned centrally by trained raters (under controlled and monitored scoring conditions) to students administered the CB format. The proposed study involves recruiting schools or districts that will agree to digitally record student responses to the Speaking tasks during the operational administration of the PB format in the first operational year. In addition to being scored live by the local test administrator for the score of record, these students' responses will be edited as

needed (so that just the responses are captured as in the CB format), uploaded to the central scoring center, and rescored by trained raters at the scoring center without the raters knowing that the performance was elicited using the PB format. As with the mode effect study, in an effort to alleviate recruiting burden, it is proposed that the study also focuses on the 1, 4-5, and 9-12 grade-level clusters. For each test form being examined in the study, about 100 students will be recruited.

The study will compare students' total Speaking raw scores assigned locally by test administrators and centrally by trained raters. Since the students take exactly the same set of Speaking tasks administered on paper, the observed differences in the scores assigned by two groups of raters can be attributed to rating condition effects (local versus central).

Writing

The Grades 4-12 Writing for the first operational year of ACCESS Online will be the same in both the PB and CB formats. These tasks had been previously administered operationally and adapted to the CB format of ACCESS Online. Given that the prompts have been previously administered, differences in student performances on the same tasks between ACCESS Online Field Test (CB format) and earlier operational administrations of ACCESS (PB format) may be primarily attributed to their administrative mode, provided that a suitable research design is employed.

Task-level Quasi Random Groups Design

A task-level quasi random groups design will be used to address the following research question: Can scores be treated as interchangeable between PB and CB formats? Should differential performances be found for a given task, its computer-based and paper-based formats will be qualitatively examined to reveal how their presentations differed and for insight into factors that may have prompted the differences in scores.

In the ACCESS Online Writing Field Test (see Table 2 for the design), all Grades 4-12 students are administered Writing prompts via computer, using tasks that had been adapted from previous operational versions of ACCESS. The Grades 4-12 students do differ, however, in how their responses were recorded: For students of both tiers in the Grades 4-5 cluster (Analysis Group II), responses were handwritten; for Tier A students in Grades 6-12 (Analysis Group III), responses were also handwritten; for Tier B/C students in Grades 6-12 (Analysis Group IV) responses were keyboarded. All students participating in the field test had also recently taken the operational ACCESS paper-based assessment in which responses were handwritten. Thus, it is only in Analysis Groups II and III that administration modes differ between operational ACCESS (PB) and ACCESS Online Field Test (CB), but response mode is constant (handwritten) – allowing the effect of administration mode to examined in isolation.

Table 2
Delivery and response mode for ACCESS Online Writing Field Test

		Tier A		Tier B/C	
Analysis Group	Grade Cluster	Delivery	Response	Delivery	Response
I	1	Paper	Handwritten	Paper	Handwritten
	2-3	Paper	Handwritten	Paper	Handwritten
II	4-5	Computer	Handwritten	Computer	Handwritten
III	6-8	Computer	Handwritten		
	9-12	Computer	Handwritten		
IV	6-8			Computer	Keyboarded
	9-12			Computer	Keyboarded

Note that students in the ACCESS Online Field Test do not take the same Writing tasks administered in the two different formats. Thus, a direct within-subject comparison of scores cannot be made using only the ACCESS Online Field Test data. Instead, the ability estimates from those students previously administered operational ACCESS will be used as a conditioning variable to create equivalent groups for each task under the two formats. Each Writing task adapted for the ACCESS Online Field Test had previously appeared on one or more administrations of operational ACCESS, yielding for each task a prior distribution of students with known ACCESS Overall Composite Scores who were administered the task on paper. Thus, for each task we have two quasi-randomly equivalent groups - one administered the task on paper and responding on paper (Task A) and a second (Analysis Groups II and III) administered the task on computer (Task A') and responding on paper – allowing for the comparison of scores at each overall ability level. The proposed design is illustrated by the dashed line in Figure 1.

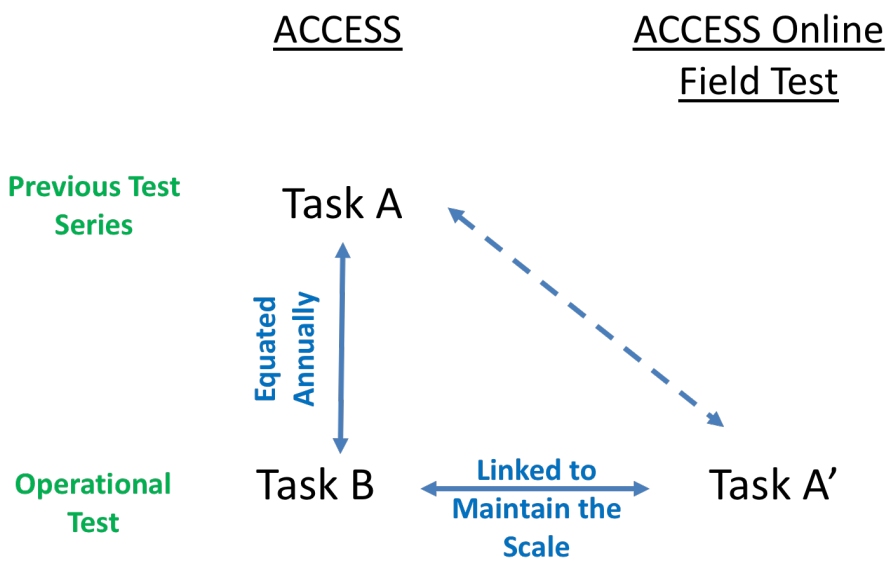

Figure 1. Task-level Quasi Random Group Design

Analysis

Responses to the ACCESS Online Writing Field Test (CB format) and ACCESS Writing (PB format) tasks included in the study are rated and transformed into Writing raw scores on the revised 0-17 scale. Please see Read-ahead A2.3 for the details of the scoring. The operational ACCESS Overall Composite score will be used to group students into various strata. For each task, the between-subject differences in administration mode will be analyzed by strata using the following procedures:

- Differences in task scores - The quasi-random groups design allows for straightforward statistical tests for differences in mean performance between groups, i.e., administration mode, at a given global proficiency level stratum.
- Differences in task score variances - An F-ratio test can be used to test for differences in score variances across the two modes of administration at each stratum.
- Differences in task score distributions - The Kolmogorov-Smirnov two-sample test of the equality of cumulative distribution functions can be used to check for equivalent score distributions at each stratum. Log-linear models or ordinal logistic regression can be used to check on the equivalency of score frequency distributions.

Limitations

The main limitation of this study is its quasi-random groups design. For each Writing task, a different matched group of students serves as the paper-based comparison group; in a proper random groups design, by contrast, two randomly equivalent groups would be assigned to an administration mode, but otherwise the analyses would take identical forms. As a compliment to this task-level study, a follow-up study will be conducted to examine mode effect for all grades using a random groups design. Details of this proposed study are presented below.

Another limitation of this study is that it does not permit the assessment of Tier B/C test takers in Grades 6-12 because of the confounding of administration mode and response mode. It should be noted, however, that operationally any Writing task presented in a paper-based format will be responded to in a handwritten response mode. It is envisioned that the results of the analyses of students from Analysis Groups II and III are generalizable to test takers in Tier B/C who may take a PB format (and produce a handwritten response) rather than a CB format (who may produce either a keyboarded or a handwritten response.) See the other read-aheads for more information on the comparability of keyboarded and handwritten responses.

Follow-Up Study: Form-Level Random Groups Design Study with Operational Data

When ACCESS Online becomes operational in 2015-2016, the comparability of the PB and the CB formats can be investigated via a random groups design since each group receives the same Writing tasks, albeit in different administrative formats. If we make the assumption that the decision to use the PB format is unrelated to the student's WIDA proficiency level, then the groups taking the PB and CB formats can be considered randomly equivalent. Thus, differences in student performances between the two formats can be attributed to mode effect. In addition, under this design, the interaction between the two modes of administration (PB and CB) and the mode of response (PB-handwritten, CB-handwritten, and CB-keyboarded) can be assessed through subgroup analyses. To help ensure that the groups being compared are truly equivalent, performances in the other domains of the operational assessment could be used as covariates.

Questions for the TAC

Question 1: Do the methods used to investigate and address score comparability within each domain (Listening, Reading, Speaking, and Writing) between CB and PB formats of ACCESS Online in its first operational year appear sound?

Question 2: What additional analyses would the TAC like to recommend to address score comparability issues between CB and PB formats of ACCESS Online in its first operational year?

References

Lottridge, S. M., Nicewander, W. A., Shulz, E. M., & Mitzel, H. C. (2010). Comparability of paper-based and computer-based tests: A review of the methodology. In C. P. Winters (Ed.), *Evaluating the comparability of scores from achievement tests variations* (pp.119-152). Washington D.C.: Council of Chief State School Officers.