Computer Testing as a Form of Accommodation for English Language Learners

Jamal Abedi
CRESST/University of California, Davis

This study compared performance of both English language learners (ELLs) and non-ELL students in Grades 4 and 8 under accommodated and nonaccommodated testing conditions. The accommodations used in this study included a computerized administration of a math test with a pop-up glossary, a customized English dictionary, extra testing time, and small-group testing. Extra time and small-group testing were included only for Grade 4 students. A reading latent composite score was used as a covariate. Results indicated that computer testing was the most effective accommodation in providing valid and accessible assessments for ELL students for both Grades 4 and 8. It is an alternative test item delivery and an easy-to-access gloss of non-math lexicon. This accommodation did not impact the validity of assessments.

Both federal and state legislation now require inclusion of all students in large-scale assessments in an effort to provide fair assessment and uphold instruction standards for every child in this country—including English language learners (ELLs). The reauthorization of Title I of the Elementary and Secondary Education Act of 1965, known as the No Child Left Behind Act of 2001 (NCLB, 2002) calls for stronger accountability and mandates inclusion of limited English proficient students and the provision of reasonable accommodations.

The challenge of serving and assessing ELL students continues to grow. According to a recent report by the U.S. Government Accountability Office (2006), about 5 million ELL students are currently enrolled in schools, representing approximately 10% of all U.S. public school students. Nationally, ELL enrollment has grown 57% since 1995, whereas the growth rate for all students has been at less than 4% (Flannery, 2009). In the state of California, approximately 1.6 million of the K–12 school population is ELLs, which makes up about one third of the nation’s ELL population (Gándara, Maxwell-Jolly, & Driscoll, 2005). This rapid growth demands that we consistently and accurately determine which students require English language services (Abedi & Gándara, 2006).

However, literature on the assessment of ELLs clearly and consistently shows a large performance gap between ELL and non-ELL students in all content areas (Abedi, 2006; Abedi...
The performance gap decreases as the level of language demand of test items decreases (Abedi, Leon, & Mirocha, 2003). Thus, the unnecessary linguistic complexity of assessments may interfere with ELL students’ ability to present a valid picture of what they know and are able to do.

To make assessments more accessible for ELL students, various accommodations are suggested in the literature and are used in the state and national assessments. However, many of the accommodations that are used for ELL students are created and used in the assessment of students with disabilities and may not be relevant for ELL students (Rivera, Collum, Willner, & Sia, 2006; Young & King, 2008). The population of students with disabilities is diverse and includes different categories of disabilities with different backgrounds and different needs. Conversely, ELL students have a common need for language assistance in order to better understand instructional materials and assessment questions. According to the NCLB legislation, accommodations can include, “to the extent practicable, assessments in the language and form most likely to yield accurate data on what [ELL] students know and can do in content areas” (see NCLB Title I, Part A, Sec. 1111 (3)(C)(ix)(II)). Research has found that many of the accommodations used in the assessment of ELL students may not be relevant for these students (Abedi, Hofstetter, & Lord, 2004). Rivera et al. indicated that ELL responsive accommodations, such as providing a dictionary and/or glossary, would be more helpful for these students. Therefore, it is imperative that the linguistic needs are considered when selecting accommodations for ELL students. However, it must be noted that not all language-based accommodations for ELL students may alter the construct being measured and therefore may not provide valid assessment outcomes for these students.

For example, two of the most commonly used accommodations for ELL students are English and bilingual dictionaries, which are considered ELL responsive accommodations. Research shows that a dictionary may provide an unfair advantage to the recipients, supplying information for the answers to the assessment questions (Abedi et al., 2004; Sireci, Li, & Scarpati, 2003). Among language-based accommodations, computer accommodations as well as the linguistic modification of assessments have been shown to be effective in reducing the performance gap between ELLs and non-ELLs and producing valid assessment outcomes.

In this study we have included several accommodations that are commonly used by states in the assessment of ELL students. These accommodations include the computerized administration of the assessment with a pop-up glossary, a customized English dictionary, extra testing time, and small-group testing. Unlike the actual dictionary accommodation, the customized dictionary does not provide unfair advantage to the recipients because all of the content related terms are removed from the customized dictionary. We examined the accommodations used in this study from three different views: (a) whether they are effective in reducing the performance gap between ELL and non-ELL students, (b) whether they produce valid assessment outcomes (they do not alter the construct being measured), and (c) whether they can be logistically applicable.

As discussed earlier, in addition to the computer-administered test, several different accommodations were used in this study including a customized English dictionary, extra testing time, and small-group testing. These accommodations were used to assess the effectiveness of computer accommodation by comparing ELL student performance under computer testing with
their performance under the other commonly used accommodations. Thus, the main focus of this study was on a computer accommodation that is capable of incorporating linguistic features to assist ELL students. We discuss computer accommodations in terms of effectiveness, validity, and feasibility. Several research questions regarding computer accommodation are addressed by this accommodation study. They include the following:

- Is computer testing as a form of accommodation effective in making content-based assessments more accessible for ELL students? (effectiveness)
- Does computer testing accommodation impact the validity of the assessment, i.e., change the content of the assessment? (validity)
- Can the computer accommodation be easily implemented? (feasibility)

METHODS

Instrumentation

Math tests were assembled for Grade 4 and Grade 8 students using a combination of National Assessment of Educational Progress and Third International Mathematics and Science Study public released items. The Grade 4 math test contained 27 questions: 1 item measuring algebra and functions, 5 in data analysis and statistics, 5 geometry, 6 measurement, 7 numbers and operations, and 3 other areas. Similarly, the Grade 8 math test contained 35 items: 5 in algebra and function, 4 in data analysis and statistics, 3 in fractions, 8 in geometry, 6 in measurement, 8 in numbers and operations, and 1 in proportionality. Reliability coefficients were computed for math tests for both Grades 4 and 8 for both forms A and B using an internal consistency (Cronbach’s alpha) approach. The alpha coefficient for Grade 4 Form A was .815 and for Form B was .812. For Grade 8 math tests, the alpha coefficient was .772 for Form A and .780 for Form B.

A reading measure is an essential part of the accommodation study since students at different levels of reading proficiency may benefit differently from the accommodations used in the study. A battery of English reading proficiency tests were chosen for this study to measure student levels of reading proficiency. The battery included the fluency subscale of Language Proficiency Scales (LAS), a National Assessment of Educational Progress (NAEP) reading block, and a test of English word recognition. Researchers from the National Center for Research on Evaluation, Standards, and Student Testing examined the content coverage of some of these English language and literacy tests (see Abedi, Courtney, Mirocha, Leon, & Goldberg, 2001; Imbens-Bailey, Dingle, & Moughamian, 1999).

A brief accommodation follow-up questionnaire was given to collect information on students’ opinions of the accommodations used in this study. Also, a questionnaire was given to collect background information from students. The background questionnaire included questions on students’ gender and ethnicity, and students’ language background, such as length of time in the United States, language other than English spoken in the home, and country of origin. A teacher and school questionnaire was also given to obtain information regarding teachers’ educational background and experience, as well as the teachers’ teaching of science and use of accommodations in the classroom. The school questionnaire included questions about school population and resources.
Accommodations Used in This Study

As indicated earlier in this article, four different accommodations were used in this study: computer accommodation with a pop-up English glossary, a customized English dictionary, extra time, and small-group testing. In addition to being used as an accommodation by itself, extra time was needed for students assigned an accommodation to make use of either the customized English dictionary or the computer test’s pop-up glossary. To absorb their extra time, students who did not receive an extra time accommodation were given a word list and asked to check off which test words were difficult to understand. The customized English dictionary accommodation consists of entries of the full dictionary without the bulk of the entire text or the unfair advantage of providing definitions for terms and concepts being tested. Students who took the computer version of the math test had access to a “pop-up glossary” (see Figure 1), a feature that provided a simple gloss of a word with the touch of the (mouse) pointer.

Small-group testing consisted of four to seven students who were tested at once in a separate room, usually a quiet school library. Under testing with extra time accommodations, eight intact Grade 4 classes were tested with extra time for completing the math test.

Sample and Design

Students in intact classrooms in Grades 4 and 8 in a single urban public school district in Southern California were tested in math and reading on 2 consecutive days. In Grade 4, 666 students from 29 classes participated, 304 of them ELL. In Grade 8, 643 students from 27 classes participated in the study, 290 of them classified as ELL. After eliminating cases with missing data, 607 students in Grade 4 and 542 students in Grade 8 were used in the subsequent analyses. The minimum number of subjects per cell was calculated through a power analysis (see Kirk, 1995, pp. 60–64) using the variance obtained in an earlier, similar National Center for Research on Evaluation, Standards, and Student Testing accommodations study (Abedi, Lord, & Hofstetter, 1998).
TABLE 1
Number of Students Tested Under Different Forms of Accommodation by ELL Categories

<table>
<thead>
<tr>
<th>Type of Accommodation/Grade</th>
<th>Total</th>
<th>ELL</th>
<th>Non-ELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>157</td>
<td>64</td>
<td>93</td>
</tr>
<tr>
<td>Small-group testing</td>
<td>20</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>Extended time</td>
<td>173</td>
<td>89</td>
<td>84</td>
</tr>
<tr>
<td>Computer testing</td>
<td>79</td>
<td>35</td>
<td>44</td>
</tr>
<tr>
<td>No accommodation</td>
<td>178</td>
<td>80</td>
<td>98</td>
</tr>
<tr>
<td>Total</td>
<td>607</td>
<td>279</td>
<td>328</td>
</tr>
<tr>
<td>Grade 8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>173</td>
<td>86</td>
<td>87</td>
</tr>
<tr>
<td>Computer testing</td>
<td>152</td>
<td>84</td>
<td>68</td>
</tr>
<tr>
<td>No accommodation</td>
<td>217</td>
<td>86</td>
<td>131</td>
</tr>
<tr>
<td>Total</td>
<td>542</td>
<td>256</td>
<td>286</td>
</tr>
</tbody>
</table>

Note. ELL = English language learner.

Grade 4 students were tested under five accommodation conditions: (a) customized English dictionary, (b) small-group testing, (c) extended time, (d) computer testing, and (e) standard testing condition with no accommodations. This generated 10 cells for Grade 4 students—5 accommodation conditions for ELL students and 5 for non-ELL students. For students in Grade 8, the accommodations were (a) customized English dictionary, (b) computer testing, and (c) no accommodation. Thus, for Grade 8, 6 cells were generated.

Of the 607 Grade 4 participants, 157 (25.9%; 64 ELL, 93 non-ELL) were assessed under the customized dictionary condition, 20 (3.3%; 11 ELL, 9 non-ELL) with small-group testing, 173 (28.5%; 89 ELL, 84 non-ELL) with extended time, and 79 students (13%; 35 ELL, 44 non-ELL) with computer testing. Testing under the standard (nonaccommodated) condition was conducted on 178 students (29.3%; 80 ELL, 98 non-ELL). The totals analyzed in each accommodation group for Grade 4 students are listed in Table 1. We used proportional random sampling techniques in assigning students to the different accommodation conditions. First we identified how many students can be tested under different accommodation conditions (logistical issues), and after identifying these allocations, we then randomly assigned students to the cells.

Of the 542 Grade 8 participants, 173 students were assessed under the customized dictionary condition (32%; 86 ELL, 87 non-ELL), 152 with computer testing (28%; 84 ELL, 68 non-ELL), and 217 under the standard (nonaccommodated) condition (40%; 86 ELL, 131 non-ELL). The totals analyzed in each accommodation group for Grade 8 students are also shown in Table 1.

Rating Open-Ended Items

The Grade 4 and 8 math and reading tests included open-ended items. NAEP scoring rubrics were used for both the math and reading tests. Raters were trained to reach the desired level of score consistency (at least an agreement of 80%). The training encouraged raters to score only the substantive content of the responses to the extent possible (rather than consider the composition, grammar, spelling, or punctuation). For 10% of the tests, two raters rated
each open-ended item separately. The interrater reliability was estimated by computing the percentage of exact and within 1 point agreement between raters, Product Moment correlation, intraclass correlation, kappa coefficient, and Williams’ index of agreement. Once interrater reliabilities were proven to be satisfactory, a single rater then scored the remainder of the items. For Grade 4 students, all items showed high interrater reliability (kappa ranging .810–.981). For Grade 4 there was more variability in the interrater reliabilities for the reading test items than the interrater reliabilities for the math test items (kappa coefficients ranging .667–.984).

All the math items for Grade 8 students showed high to perfect interrater consistency (kappa ranging .854–.990). The interrater reliabilities for the reading test items for Grade 8 students were lower (kappa ranging .708–.942) than for the math test items, with one item posing considerable difficulty for the raters ($\kappa = .529$).

**Rating of Items for Linguistic Complexity**

Our previous studies have clearly shown that linguistic complexity of content-based test items negatively impacts the performance of ELL students (see, e.g., Abedi, 2006; Abedi, Courtney, & Leon, 2001; Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997; Wolf et al., 2008). To examine such an effect, and to control for linguistic complexity of math test items, individual test items were rated for linguistic complexity.

The linguistic complexity rating was based on the rubric developed in our earlier studies (Abedi, 2006; Abedi et al., 2001). These ratings are composites from the scores given by two college instructors of English grammar. A 5-point Likert scale was used for rating linguistic difficulty of the items ranging 0 (less linguistically complex) to 4 (very linguistically complex). Items were grouped into two categories based on their level of linguistic complexity, less complex (0, 1, or 2) and more complex (3 and 4). Two testlets were created accordingly. Hypotheses of effectiveness and validity of accommodations were examined separately for each of the two testlets.

In this study a latent composite of multiple measures of students’ English proficiency was used as a covariate. The reading latent variable was defined as the common variance among three measures: LAS fluency section, NAEP open-ended reading comprehension questions, and NAEP multiple-choice reading questions. To be used as an effective covariate, a variable must be highly correlated with the outcome variable. The correlation between the two latent composites (reading and math) was .559, which is large enough to justify the use of reading composite as a covariate for math scores.

**RESULTS**

Three research questions guided design, data collection, data analyses, and reporting in this study. These research questions are as follows:

- Is the computer testing as a form of accommodation effective in making content-based assessments more accessible for ELL students? (effectiveness)
- Does computer testing accommodation impact the validity of the assessment, that is, change the content of the assessment? (validity)
- Can the computer accommodation be easily implemented? (feasibility)

Results for Grade 4 Students

**Effectiveness.** ELL and non-ELL students in Grade 4 were tested under four different accommodations and a standard condition. Students tested under the standard condition served as a control group. Non-ELL students were tested under the five testing conditions to provide data for examining the validity of accommodations used in this study. A comparison of the math scores of ELL students receiving different forms of accommodation with the scores of ELL students under the standard testing condition provided evidence on the effectiveness of the accommodations for ELL students. Those accommodations that helped ELL students perform significantly higher than their peers testing under the standard (no accommodation) condition were labeled as effective in making assessments more accessible for ELL students. The higher the difference between the accommodated and nonaccommodated performance the more effective the accommodation was.

Table 2 presents the means, standard deviations, and numbers of students for each of the five conditions (four accommodations and the standard condition) for Grade 4 ELL students. As the data in Table 2 show, 279 students were tested under the five conditions. Because there were limits to how many students could be tested simultaneously on computers, a smaller number of students were tested under the computer accommodation condition \((n = 35)\). The mean score for students under the computer accommodation was 14.69 \((SD = 5.12, n = 35)\); under the extra time accommodation, the mean was 13.74 \((SD = 6.02, n = 89)\); under the customized dictionary accommodation, the mean was 13.81 \((SD = 6.04, n = 64)\); under the small-group accommodation, the mean was 9.55 \((SD = 2.88, n = 11)\); and under the standard condition the mean math score was 12.27 \((SD = 5.24, n = 80)\). As these data suggest, students in Grade 4 performed better under most forms of accommodation that were provided in this study. However, the data also clearly indicate that students benefited more from some forms of accommodation than others. For example, students taking the computer test obtained 2.42 score points (about .5 \(SD\)) higher than students under the standard condition. This difference becomes smaller with the other accommodations. The difference in student performance under

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>(N)</th>
<th>(M)</th>
<th>(SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>35</td>
<td>14.69</td>
<td>5.12</td>
</tr>
<tr>
<td>Extra time</td>
<td>89</td>
<td>13.74</td>
<td>6.02</td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>64</td>
<td>13.81</td>
<td>6.04</td>
</tr>
<tr>
<td>Small-group testing</td>
<td>11</td>
<td>9.55</td>
<td>2.88</td>
</tr>
<tr>
<td>Standard condition</td>
<td>80</td>
<td>12.27</td>
<td>5.24</td>
</tr>
<tr>
<td>Total</td>
<td>279</td>
<td>13.44</td>
<td>5.72</td>
</tr>
</tbody>
</table>

*Note.* ELL = English language learner.
extra time and standard condition was 1.47 (about .25 $SD$), for the customized dictionary the
difference was 1.54 (about .25 $SD$), and for small-group testing the difference was $-2.72$
(students under small-group testing performed on average 2.72 lower than those under the
standard testing condition).

The data in Table 2 suggest that students performed better under all accommodation condi-
tions except small-group testing. Because the number of students who participated in small-
group testing was very small, descriptive statistics for this accommodation may not be reliable;
therefore, further analyses were not performed on the data from this accommodation.

These accommodations are considered effective if the difference in performance under
accommodated versus non-accommodated conditions is statistically significant. We randomly
assigned students to the different accommodation conditions (eight groups: four ELL, four non-
ELL). However, in spite of randomization of students to the eight accommodation conditions,
initial English proficiency differences may have existed because of small sample size in some
groups. To adjust for any initial difference in the level of English proficiency, the reading
score, a latent composite of the reading components (LAS subscale score plus multiple-choice
and open-ended NAEP reading scores), was used as a covariate. The differences among
the adjusted means under the accommodated and the standard conditions were tested for
statistical significance using student’s multiple $t$ test as a set of a priori comparisons adjusted
for the Type I error rate (Kirk, 1995, pp. 129–134). Table 3 presents the results of these
comparisons.

The half standard deviation difference (between computer testing and the standard condition)
in Table 2 was tested for statistical significance. The results in Table 3 show this difference
was significant at the .005 nominal level, well beyond the .01 nominal level. To show the
magnitude of such effects, we also computed the coefficient of determination, $\eta^2$, the percentage
of variance of math explained by the accommodations. For the effectiveness of the computer
testing condition, $\eta^2$ was .030, suggesting that this accommodation affects math performance
of about 3% (explaining 3% of the variance of the math score), which may not be considered
as a major impact.

For extra time, the difference between accommodated and nonaccommodated ELL students
was also significant at the .01 nominal level. For this comparison, $\eta^2$ was .024, which explains
about 2.4% of the variance of math scores. For the customized dictionary, the difference between
the ELL students’ accommodated and nonaccommodated performance was not statistically

<table>
<thead>
<tr>
<th>TABLE 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grade 4 Math Means for ELL Students Adjusted by Reading Scores</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>$N$</th>
<th>Adj. $M$</th>
<th>$SE$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>35</td>
<td>14.922</td>
<td>.805</td>
<td>.005</td>
</tr>
<tr>
<td>Extra time</td>
<td>89</td>
<td>14.037</td>
<td>.506</td>
<td>.012</td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>64</td>
<td>13.372</td>
<td>.597</td>
<td>.138</td>
</tr>
<tr>
<td>Small-group testing</td>
<td>11 (NA)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Standard condition</td>
<td>80</td>
<td>12.182</td>
<td>.533</td>
<td></td>
</tr>
</tbody>
</table>

Note. Each student’s reading score was used as a covariate. ELL = English language learner; NA = not applicable.
significant. Small-group accommodation results were not tested for statistical significance because of the small sample size. The results of analyses for students in Grade 4 suggest that computer testing and testing with extra time were effective forms of accommodation for ELL students.

**Validity.** To consider an accommodation strategy for use in national and state assessments, it must be effective and produce valid assessment outcomes. An accommodation is valid if it does not change the construct under measurement, that is, if it does not increase the performance of non-ELL students, or if it affects performance of ELLs more than performance of non-ELLS. To test the validity of accommodations, non-ELL students were also tested under the five different accommodation conditions. Table 4 presents descriptive statistics, including the mean, standard deviation, and number of students for each of the five accommodation conditions for non-ELL students. Comparing the mean math score under the accommodations with the mean under the standard testing condition, one may not see any improvement in students’ performance due to the use of accommodations. For the computer testing, the mean math for non-ELL students was 16.45 (SD = 5.63, n = 44); for extra time, the mean was 16.47 (SD = 6.30, n = 84); for the customized dictionary, the mean was 17.46 (SD = 7.03, n = 93); and for small-group testing, the mean was 15.56 (SD = 6.56, n = 9), compared to a mean of 17.38 (SD = 6.95, n = 98) for the standard (nonaccommodated) condition.

To control for any initial differences between groups, we adjusted the mean math score by the student’s level of English proficiency using a reading score composite as a covariate. Although the differences between the accommodated and standard (nonaccommodated) conditions did not seem to be large, we tested these differences for statistical significance. Table 5 shows the results of analyses comparing non-ELL, accommodated performance with non-ELL, nonaccommodated performance. To control for the Type I error rate, the pooled-within-group variance was used in a set of a planned (a priori) multiple t tests. These analyses were performed on the adjusted scores.

As the data in Table 5 suggest, none of the differences were statistically significant. That is, non-ELL students performed the same under the accommodated and nonaccommodated (standard testing) conditions. These results suggest that the accommodations that were found to be effective for students in Grade 4 were also valid because they did not affect the construct (math performance).

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>44</td>
<td>16.45</td>
<td>5.63</td>
</tr>
<tr>
<td>Extra time</td>
<td>84</td>
<td>16.47</td>
<td>6.30</td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>93</td>
<td>17.46</td>
<td>7.03</td>
</tr>
<tr>
<td>Small-group testing</td>
<td>9</td>
<td>15.56</td>
<td>6.56</td>
</tr>
<tr>
<td>Standard condition</td>
<td>98</td>
<td>17.38</td>
<td>6.95</td>
</tr>
<tr>
<td>Total</td>
<td>328</td>
<td>17.00</td>
<td>6.62</td>
</tr>
</tbody>
</table>

*Note. ELL = English language learner.*
Feasibility. The feasibility of each accommodation was determined by the test administrators’ observations and the project staff’s experience. The use of computer versions of the math tests required programming of the tests with the pop-up glossaries, as well as administering the tests with computer equipment and access to the Internet. With the computer testing, the main obstacle was school access to the proper Internet wiring, computer memory, and Internet software. In some schools, there was not a quiet place for a group of students to take a computer test. Another consideration was student agility with aiming the mouse and with moving the scroll bar. Some students closed the test window accidentally.

The greatest logistical issue for the extra time accommodation was scheduling. Schools were reluctant to allow extra test administration time that borrowed from teaching time. A second challenge was keeping the test environment quiet when there were Grade 4 students who had turned in their tests before the extra time was over.

The lack of use of the customized English dictionary in the paper-and-pencil version of the assessment led us to look for evidence of lack of exposure to dictionary use. We found in the student background questionnaire that only 12 of the Grade 4 students (7.6%) provided with the customized English dictionary had used an English dictionary in the classroom before (see Table 6).

Looking at the responses of all the Grade 4 participants, only 55 students (9.1%) responded that they had used a dictionary in the classroom before. Similarly, when asked on the accommodation follow-up questionnaire which accommodations they would prefer, only 26 (16.6%)

### TABLE 5
Grade 4 Math Means for Non-ELL Students—Adjusted by Reading Scores

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>N</th>
<th>Adj. M</th>
<th>SE</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>44</td>
<td>16.295</td>
<td>.822</td>
<td>.262</td>
</tr>
<tr>
<td>Extra time</td>
<td>84</td>
<td>16.550</td>
<td>.595</td>
<td>.292</td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>93</td>
<td>17.435</td>
<td>.565</td>
<td>.971</td>
</tr>
<tr>
<td>Small-group testing</td>
<td>9</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Standard condition</td>
<td>98</td>
<td>17.406</td>
<td>.550</td>
<td></td>
</tr>
</tbody>
</table>

*Note. Each student’s reading score was used as a covariate. ELL = English language learner.*
of the students provided with the customized English dictionary and 96 of the total Grade 4 students (15.8%) stated that they would like an English dictionary “to make it easier for [them] to understand math problems” (see Table 7).

Results for Grade 8 Students

**Effectiveness.** Grade 8 students were assessed under the computer testing and customized dictionary accommodations as well as under the standard condition (no accommodation provided). Table 8 presents descriptive statistics including mean, standard deviation, and number of ELL students for the accommodated and nonaccommodated conditions. As the data in Table 8 show, ELL students performed higher under both accommodations when compared to the performance under the standard condition. For the computer testing, the mean was 10.17 (SD = 4.36, n = 84); for the customized dictionary, it was 9.95 (SD = 3.84, n = 86); and for the standard condition, the mean was 9.47 (SD = 4.00, n = 86).

Similar to the analytical approach used for Grade 4 data, we used a reading score composite as a covariate to control for possible initial differences among the 6 groups. A series of planned (a priori) comparisons were conducted on the adjusted means between student performance under the accommodated conditions and the performance under the standard testing condition. Table 9 presents the results of these comparisons. As the data in Table 9 show, increased performance of ELL students under computer testing was significant beyond the .01 nominal
TABLE 9
Grade 8 Math Total Means for ELL Students—Adjusted by Reading Scores

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>N</th>
<th>Adj. M</th>
<th>SE</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>84</td>
<td>10.66</td>
<td>.408</td>
<td>.008</td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>86</td>
<td>9.84</td>
<td>.399</td>
<td>.197</td>
</tr>
<tr>
<td>Standard condition</td>
<td>86</td>
<td>9.11</td>
<td>.401</td>
<td></td>
</tr>
</tbody>
</table>

Note. Each student’s reading score was used as a covariate. ELL = English language learner.

level \((p = .008)\). However, the increased performance under the customized dictionary condition was not significant for these students \((p = .197)\). This suggests that computer testing was an effective accommodation for ELL students in Grade 8 as well.

Validity. Table 10 presents descriptive statistics (mean, standard deviation, and number of students per group) for Grade 8 non-ELL students under the three accommodation conditions (computer testing, customized dictionary, and standard condition). As the data in Table 10 suggest, the means for the computer testing and for the customized dictionary are slightly higher than the mean math under the standard condition. For the computer testing, the mean was 14.76 \((SD = 4.55, n = 68)\); for the customized dictionary, the mean was 14.12 \((SD = 4.28, n = 87)\), compared to a mean of 13.94 \((SD = 4.57, n = 131)\) for students tested under the standard condition.

Table 11 presents the results of planned multiple comparisons for non-ELL students. The adjusted mean math scores of non-ELL students under each accommodation were compared to the adjusted mean scores of non-ELL students under the standard testing condition. As the

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>N</th>
<th>Adj. M</th>
<th>SE</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>68</td>
<td>14.674</td>
<td>.491</td>
<td>.220</td>
</tr>
<tr>
<td>Customized dictionary</td>
<td>87</td>
<td>14.205</td>
<td>.434</td>
<td>.623</td>
</tr>
<tr>
<td>Standard condition</td>
<td>131</td>
<td>13.930</td>
<td>.354</td>
<td></td>
</tr>
</tbody>
</table>

Note. Each student’s reading score was used as a covariate. ELL = English language learner.
data in Table 11 show, none of the comparisons were significant. That is, the two accommodation strategies did not affect the performance of non-ELL students. This suggests that the accommodations used in this study can be implemented without concerns about their validity.

Feasibility. The feasibility of administering tests with a customized glossary and computer accommodations is possibly greater for Grade 8 students. Feasibility more likely increases with Grade 8 students because their exposure to dictionary and computer use may be greater and their schools have more computers and more Internet access connections.

Accommodation Impact on Assessments With Different Levels of Linguistic Complexity

Individual test items were rated for linguistic complexity. The linguistic complexity rating was based on the rubric developed in our earlier studies (Abedi, Courtney, & Leon, 2002). A 5-point Likert scale was used for rating linguistic difficulty of the items ranging 0 (less linguistically complex) to 4 (very linguistically complex). We combined items into two categories, less complex (rated 0, 1, or 2) and more complex (rated 3 or 4) and created two testlets accordingly. We examined the hypotheses of effectiveness and validity separately for each of the two testlets.

The effects of accommodations across the two testlets by varying degree of linguistic complexity were examined using a multivariate analysis of covariance model. The results indicated that for Grade 4 ELL students, all the accommodations made a significant difference for the more linguistically complex items (computer, \( p = .017 \); extra time, \( p = .027 \); customized dictionary, \( p = .049 \)). For the less complex math items, the computer and extra time accommodations were still significant. For non-ELL students, there were no significant accommodation effects; therefore, there were no validity issues for either of the two testlets.

For Grade 8 ELL students, the computer accommodation produced significant results for the more linguistically complex items (\( p = .001 \)), but it was not significant for the items that were less linguistically complex. For non-ELL Grade 8 students, there was no significant accommodation effect; therefore, validity was not a concern for either testlet.

Very few students marked circles to indicate that they had looked up words in the customized English dictionary. In Grade 8 classes, 140 of the 204 students with customized dictionaries marked the sample word they were asked to find. Otherwise a maximum of 4 students marked any given word, such as growth, on the pages of definitions. In Grade 4 classes, 146 of the 170 students with customized dictionaries marked the sample word as instructed. Technical words such as grid, width, and length were looked up and marked by 8, 7, and 5 students, respectively.

The lack of marked circles confirmed the test administrators’ observations that most students did not use the customized dictionary or word lists after the first few attempts to look up a math word. In looking for evidence for lack of exposure to dictionary use, we found in the student background questionnaire that only 26 (15.0%) of the Grade 8 students provided with the customized English dictionary stated that they had used “an English dictionary” in the classroom before. Of all the Grade 8 participants, only 78 students (14.4%) responded that they had used a dictionary in the classroom before. Similarly, when asked on the accommodation follow-up questionnaire which accommodations they would prefer, only 22 (12.7%) of the students provided with the customized English dictionary and 80 of the total Grade 8 students
(14.8%) stated that they would like an English dictionary “to make it easier for me to understand math problems.”

Students taking the computer version of the math test had access to a “pop-up glossary,” a feature that provided a simple gloss of words when students pointed to them with the mouse. The program timed the length of time students spent on each test item and the time that the gloss appeared on the screen. The results in Grade 8 produced a larger difference between the glossing behavior of ELL and non-ELL students. ELL students in Grade 8 spent nearly three times as much time glossing, and glossed almost twice as many words as non-ELL students.

Grade 8 students who took the computer test were much more likely to report that taking the test was “fun” than students with the customized dictionary or standard condition. These students were also more likely to want “more tests like this one.” A higher percentage of students in Grade 8 who took the computer- and customized-dictionary-accommodated tests reported that the test was easy compared to those who took the standard condition test. Similarly, a lower percentage of those accommodated felt that the math test was hard. Students in Grade 8 who took the computer accommodation reported looking up more words during the test than those using the customized dictionary.

In summary, students who were tested under the computer accommodation (particularly the ELLs) indicated that testing under this accommodation was easy and more fun. This accommodation also makes all students (particularly ELLs) more inclined to look up more words than they would have with the customized English dictionary.

DISCUSSION

To include ELLs in the state and national assessments and to provide a fair and valid assessment for them, some forms of assessment accommodations have been provided. The main objective of this study was to identify accommodations that can help ELL students with their linguistic needs without altering the construct being measured. To identify effective and valid accommodations for these students, two sets of accommodation strategies were included in this study. The first set was selected from those used in the NAEP and were found to be effective in increasing the inclusion rate. The second set of accommodations was language related and was among those that researchers found to be effective in reducing the performance-gap between ELL students and non-ELL students.

The results of our analyses for Grade 4 revealed that extra time and computer testing were effective forms of accommodation for ELL students. For non-ELL students, these two accommodations did not affect their assessment outcomes. Therefore, some of the accommodation strategies used in this study showed effectiveness in making assessments more accessible for Grade 4 ELL students without posing any threat to the validity of the assessment. The results also indicate that only computer testing is an effective accommodation for the Grade 8 ELL students in this study. This accommodation has no impact on the assessment of non-ELL students, suggesting that the computer testing for Grade 8 can be implemented without a validity concern.

One of the most important characteristics of the computer testing was the extensive use of its pop-up glossary by the students. Under the customized English dictionary accommodation, students rarely indicated that they used the customized dictionary. The computer program
recorded the length of time students spent on each test item. The computer also kept track of which glossary items students looked up and for how long the mouse stayed in that position. The results indicated a large difference between the glossing behavior of ELL and non-ELL students. For example, ELL students in Grade 8 spent nearly three times as much time glossing, and glossed almost twice as many words as non-ELL students.

Delivery of the customized dictionary by computer had several advantages for the students. Students pointed the mouse to an unknown word instead of searching for it in an alphabetical collection. Students were presented with the dictionary entry of only that word (or its root) in its present context, rather than being given all the possible definition entries.

Taking a test on a computer—usually in a special room and in a group of about eight—may have been perceived as a privilege rather than chore. We expected that the randomly selected non-ELL students would also perform better than their “paper-test” peers, but the Grade 4 students did not, and the slight increase in the performance of Grade 8 students did not reach a significant level. The slight difference may be accounted for by familiarity with the keyboard and mouse. Another consideration from test administrator observation is that more non-ELL students possessed the touch-typing skills that made responding to open-ended questions faster. Students expressed enjoyment of the computer delivery of the test, despite the predominance of “hunt and peck” typing. As discussed in the Results section of this article, all students indicated in their background questionnaires that they had more fun with computer testing than with any other accommodation used in this study.

Because Internet access was required for administering the computer version of the math and reading tests, testing was limited to certain schools, certain rooms, and computers of a certain size. When we needed to bring laptops to the site, there was at least an hour of setup and another of clean up. We sometimes had to borrow some of the static IP addresses for that school and type a unique IP address into each laptop’s Internet set-up menu the day before. When using a school’s Macintosh computers, some browser versions could not display the pop-up glossary properly, so we had to load new browsers.

Accommodation Impact on Measurement With Degrees of Linguistic Complexity

As indicated earlier, math test items were categorized into two categories based on the level of their linguistic complexity (less complex with a linguistic rating of 0, 1, or 2, and more complex with ratings of 3 and 4). For the more linguistically complex items, all the accommodations made a significant difference for Grade 4 ELL students. For Grade 8 ELL students, we found that the computer accommodation was significant for the more linguistically complex items ($p = .001$), but it was not significant for the items that were less linguistically complex. This is additional evidence of the validity of the accommodations, illustrating that the computer accommodation for Grade 8 students only showed an effect on those items for which we would expect language to most disadvantage ELLs. For the less complex Grade 4 math items, the computer and extra time accommodations were still significant for ELL students. For non-ELL students in Grades 4 and 8, there was no significant accommodation effect; therefore, validity was not a concern for either testlet in either grade.

In this study, we find that the computer accommodation is effective and valid. That is, it can be used on both ELL and non-ELL students without the concern of changing the focal
construct. Thus, we recommend this accommodation for ELL students. This use, of course, is dependent on an increase in the feasibility of putting together computer tests and administering them at school sites.

ACKNOWLEDGMENTS

The work reported here was supported in part under a grant from the U.S. Department of Education, Institute for Education Sciences. The findings and opinions expressed here do not necessarily reflect the position or policies of the Institute for Education Sciences or the U.S. Department of Education. I acknowledge the contribution of several people. Seth Leon contributed to this article by helping with statistical analyses of the data and providing feedback and suggestions on the design of the analyses. Mary Courtney helped with the literature review and with the writing of the initial report of this study. Kira Luna assisted in structuring and revising the article.

REFERENCES


