Study of a Dual-Language Test Booklet in Eighth-Grade Mathematics

Teresa García Duncan
*American Institutes for Research*
*Washington, DC*

Lourdes del Río Parent
*Foothill-De Anza Community College District*
*Los Altos Hills, California*

Wen-Hung Chen, Steve Ferrara, Eugene Johnson, Scott Oppler, and Yann-Yann Shieh
*American Institutes for Research*
*Washington, DC*

The purpose of this study was to address the effectiveness of a Spanish–English dual-language test booklet in 8th-grade mathematics. This study used analyses of test data (n = 402) as well as information gathered through focus groups (n = 68) and cognitive interviews (n = 18) to assess the accuracy and utility of the dual-language format as a testing accommodation. The data indicate that the translation procedures were effective and that the 2 test booklet formats were comparable once English-language proficiency and language used to answer the test were accounted for. Students found the dual-language format to be very useful, and no undue advantage appeared to be accorded to students given the dual-language booklet over the English-only booklet. In our discussion, we address translation issues and the challenges in accurately identifying students who should receive language accommodations.

Inclusiveness is a key issue in large-scale assessments. For English-language learners, testing accommodations include tests translated into students’ native languages and provision of additional time or bilingual dictionaries. The purpose of
this study was to address the effectiveness of a Spanish–English, dual-language (DL) test booklet in eighth-grade mathematics. This accommodation involves a booklet format in which the original English items are placed on one side of the booklet and the corresponding items translated into Spanish placed onto the facing pages. We conducted analyses of students’ performance, focus groups, and cognitive interviews (think alouds) to determine whether the DL and English-only versions of this mathematics test were comparable to one another.

We conducted this study to assist the National Assessment Governing Board in developing inclusion and accommodation policies for the previously proposed Voluntary National Test (VNT). An accommodation for limited English proficient (LEP) students that was being considered was a DL test booklet, and this study was directed toward answering a series of questions regarding the equivalency, quality, and usability of a DL test booklet compared to an English-only booklet format. Although the VNT program is now defunct, the research questions explored in this study have bearing on other large-scale assessments. Several subtasks comprised the DL test booklet study:

- Study 1: Evaluation of the comparability of the DL and English-only booklets via traditional analyses (e.g., differential item functioning [DIF]; adjusted and nonadjusted mean score differences).
- Study 2: Focus groups of students conducted immediately after they took the mathematics test to document students’ overall experience with the two types of booklets.
- Study 3: Cognitive interviews to obtain in-depth information on the validity of the translation and about how students used the DL test.

The three studies that comprise this body of research allow for an investigation into the cognitive processes that bilingual and Spanish-speaking LEP students employ when using the DL test booklet. In addition, these studies provide information about factors other than mathematical knowledge and problem-solving ability that may have an effect on students’ test performance. The studies offer answers to the following research questions:

- Is the content of the DL version of the test questions comparable to the English version?
- Do students understand the DL version of the test questions as a vehicle for assessing mathematics?
- Is the DL version clear and acceptable to the various communities in the United States for whom Spanish is the native language?
- Are the grammar and language structure used in the DL version correct?
METHOD

Study Design

We used National Assessment of Educational Progress (NAEP) inclusion criteria as the basis for identifying students who would qualify for an accommodation such as the DL test booklet (Anderson, Jenkins, & Miller, 1996). Indeed, similar criteria exist for other large-scale assessments, such as those administered by states (National Center for Education Statistics [NCES], 1997; Olson & Goldstein, 1997; Rivera, Stansfield, Scialdone, & Sharkey, 2000). Currently, students who have had fewer than 3 years of academic instruction in English are excluded from NAEP administrations unless one of the following two conditions are met: (a) the NAEP test is available in the student’s native language (currently, Spanish is the only language other than English in which the NAEP is administered) or (b) school administrators judge that the student is capable of participating in the assessment in English.

Therefore, one of the examinee groups for this study was composed of native Spanish-speaking students who had fewer than 3 years of academic instruction in English and who would not be able to participate in a NAEP assessment in an English-only format. In other words, these are the students who would be targeted for an accommodation such as a DL test booklet. However, to evaluate fully the effectiveness and utility of a DL test booklet, it was necessary to add into the design native Spanish-speaking students who have had greater than 3 years of academic instruction in English. Although these students would be not be offered any language-related accommodations in NAEP administrations, the clearest way to evaluate the booklet effect would be to assign randomly into the two booklet conditions individuals who could use either a DL or an English-only test booklet and compare their performances. Native English speakers were also included in the study design as a control group.

Recruitment

Sampling procedures. To allow for sufficient statistical power, the target sample size for this study was initially proposed to be 1,600 (400 students per cell). A probability sample of 50 schools was planned, with data collection to occur in late spring 2000. Unfortunately, the timing of the data collection forced us to abandon the probability sample and indeed required us to simply use as many schools as were willing to participate in the study. The Council of the Great City Schools (CGCS) led our recruitment efforts. Given the relevance of DL assessments to the CGCS constituents, we focused recruitment on CGCS member school districts.
Our recruitment efforts, combined with an incentive package valued at approximately $325, yielded a total of 10 schools and a sample size of 402.

**Characteristics of schools and students sampled.** The 10 participating schools were all high-poverty schools with large minority student populations. These were appropriate for this study in the sense of having large populations of the Latino students we sought. With regard to the student groups we were targeting in our study, we sampled equal numbers of each of the four groups (Groups A–D; see Table 1) in each school. However, not all students sampled appeared for the test administration. Participation rates within each of the four study groups are shown in Table 1. Although it is unclear why participation rates for Groups B and D are lower than those in Groups A and C, the participation rate for students assigned into the DL test booklet condition is the same as for students assigned into the English-only test booklet condition (69.79%).

Each incentive package contained a collection of Spanish–English dictionaries, copies of the study’s research report, copies of *Study of a Dual Language Test Booklet in 8th Grade Mathematics: Answer Key and Scoring Guide* (American Institutes for Research, 2000), 30 subscriptions to an educational magazine, and educational software. Schools received multiples of this set of materials corresponding to every 30 students recruited from their buildings.
Test Assembly

The mathematics achievement test used for this study was assembled using 60 NAEP items but followed VNT mathematics (VNT–M) test specifications. The pool of items from which the 60 test questions were chosen consisted of the 1990, 1992, and 1996 NAEP eighth-grade mathematics item banks. Two experienced staff members with mathematics content expertise assembled the test.

The first pass through the item pools involved identifying the test questions that seemed to be the most clearly written and direct in conveying the intent of the question. This clarity criterion was implemented for two reasons: (a) to facilitate translation of the item from English to Spanish and (b) to ensure to the extent possible that the assembled test was a test of mathematics and not of reading ability. After winnowing out the items that met this criterion, we then began constructing a test according to VNT–M test specifications (American Institutes for Research, 1999; National Assessment Governing Board, 1998). Because three different item pools were used (1990, 1992, and 1996), the items were first equated by using the linear transformation constants provided in the NAEP technical reports, then the test information function (TIF) was calculated. To attain the desired TIF, an iterative procedure was initiated: Using the item statistics provided by NCES (item p values and item response theory parameters), items were removed and replaced and the TIF recalculated until we obtained a TIF as similar to the target function as possible. Although our intent was to use public release NAEP items only, it became necessary to use secure items to obtain the TIF we sought (consequently, 4 of the 60 test items in the test form assembled were secure NAEP items). Although we sought more of a plateau between the basic and proficient levels, the obtained TIF met our purposes given the constraints of (a) being able to use only NAEP items instead of VNT items and (b) using as limited a number of secure NAEP items as possible.

Translation Procedures

Selecting items and meeting VNT–M test specifications gave us an English-only test booklet; the next undertaking was creating a DL test booklet. In light of the research and recommendations reviewed in the test adaptation literature, we identified a translation team whose members (a) represented and were familiar with different Latino cultures and (b) had expertise in mathematics assessments. The translation team was composed of six American Institutes for Research staff members and two expert consultants.

The translation proceeded according to the following steps.

**Step 1: Forward translation of original English booklet (Version A) to Spanish (Version B).** Translator 1 did the source-to-target language translation.
Step 2: Back translation of Spanish booklet (Version B) to English (Version C). Translator 2 did the target-to-source language translation.

Step 3: Examination of the equivalency of the back-translated document (Version C) to the original English booklet (Version A). Native English speakers evaluated the similarity of meaning between Version A and Version C. An evaluation form was used to document any problems that were identified. Translators 1 and 2 met to resolve the problems identified at this stage. The Microsoft® Word document statistics function was used to estimate differences in (a) word and sentence complexity and (b) word and sentence length.

Step 4: Examination of the equivalency of the back translation (Version C) to the original English booklet (Version A) and to the forward translation (Version B). To ensure that the two translators used the same translation rules and that the back-translated document (Version C) did not mask or hide problems in the original translation, bilinguals evaluated all three versions of the test. Evaluation forms were used to document any problems with words or phrases that sounded stilted or unnatural. The committee met to resolve the difficulties identified.

Step 5: Evaluation of the Spanish translation (Version B) with respect to clarity, length of words and sentences, and reading level. Four bilingual judges reviewed the Spanish version for clarity, existence of stilted phrases, and appropriateness of reading level. An evaluation form was used to document any problems that were identified. These judges then met with the two translators to resolve the difficulties. The Microsoft Word document statistics function was used to assess word and sentence length.

We should also note that the translation team made two important decisions regarding the test adaptation. First, the translation team agreed that it would be most prudent to use the formal usted form instead of the familiar tu. In certain Latino cultures, one uses the tu form only with family members and close friends; being addressed in the familiar by others can be offensive. To avoid this problem, we used the usted form—which may sound very formal to some but is universally considered the polite form of address—in the translated test. The second decision had to do with using words that would be familiar across different Latino groups. There are many Spanish dialects, and in the cases in which the committee could not agree on a single most appropriate translation, a second word was inserted in parentheses to ensure that all Spanish-speaking students would understand the question. For example, arrendar and alquilar are two ways of expressing the verb to rent: “El costo de arrendar (alquilar) una motocicleta se obtiene con la siguiente fórmula.”

The DL booklet listed the Spanish versions of the items on the left-hand pages and the English versions of the items on the right-hand pages. Please note that the translation and quality control procedures used here are necessary but not suffi-
cient conditions for establishing the accuracy and quality of the test adaptation. Empirical data must also be used to gauge the quality of the adaptation. Our quantitative and qualitative analyses address this issue; we discuss those results in subsequent sections of this article.

Language Background Questionnaire

In addition to the two sections of the test, the final section of the test booklet consisted of 23 items asking students about their language background (e.g., race, ethnicity, years in United States, self-ratings of language proficiency). The DL test booklet also contained two additional questions regarding how the students used the test booklet and how useful they found it to be. The language background questionnaire was adapted from the work of Abedi and his colleagues (e.g., Abedi, Lord, & Hofstetter, 1998; Abedi, Lord, & Plummer, 1997). The data from this questionnaire were included for use as possible covariates in our analyses as well as for documentation of the demographic and language ability characteristics of our sample.

Test Administration

A team of experienced Westat field supervisors conducted the data collection. All of the eight field staff members were veteran NAEP administrators, and several also had experience with the Trends in International Mathematics and Science Study and the Program for International Student Assessment. Three of the administrators were bilingual (two were native Spanish speakers of Mexican descent). The bilingual supervisors led the DL administrations and the focus groups, whereas the other field staff led the English-only test sessions. The test administration protocol developed for the DL study was an adaptation of the NAEP script. We developed a bilingual script for the DL administrations and a parallel English-only script for the English administrations.

Computing Test Performance

The test assembled consisted of 60 items (45 multiple choice and 15 constructed response). Incorrect responses were always scored zero, as were double-gridded responses. Only eight students (five of whom were from Group A and three from Group B) used both sides of the DL booklet and replied twice to the same question. Any discrepancy between the Spanish answer and the English answer was treated as a double-gridded response and marked as incorrect. Blank, crossed out, illegible, and off-topic responses were treated as missing data. The maximum score for the assembled test was 77 points. Each multiple-choice item was worth 1 point, and the constructed-response items ranged from 1 to 4 points.
Scoring Constructed-Response Items

Our colleagues at Pearson-NCS took the lead in selecting and training scorers. The 2000 NAEP assessment used bilingual booklets in fourth- and eighth-grade mathematics tests, and so scorers for the DL study were recruited from the pool of staff members who performed that scoring. A total of six bilingual raters scored the constructed-response data from this study.

**Interrater reliability.** All constructed-response items were double scored to obtain a detailed portrait of the types of items (in English or in Spanish) that might cause difficulties in rater agreement. Interrater agreement across the two extended and the 13 short constructed-response items was excellent. Using Cohen’s kappa and Pearson’s $r$ as our measures of agreement, we found that for the set of Spanish-language responses, the average kappa was .91, and the average $r$ was .94. The values for the set of English-language responses were exactly the same as the Spanish set. When all pairs of scores were pooled across the two languages, the average kappa was .92 and the average $r$ was .95. Of the 5,408 responses coded, a total of only 215 disagreements (3.97%) were observed. Of those 215 disagreements, only 37 (or 0.68% of all coded responses) were nonadjacent (i.e., those greater than 1 score point apart). The high consistency may be due to the quality of the training, the experienced coders, and the relatively small number of test booklets coded during a short period of time (1 month). Because reliability was so high, we chose to use the first score of the pair to compute student test performance.

**STUDY 1: TEST PERFORMANCE ANALYSES**

As shown in Table 2, our study followed a balanced, mixed model design with two between-group factors (NAEP inclusion status and booklet type) and one within-group factor (standard vs. extended time). With this design, we sought to address the following:

- The speededness of the test by examining student reached and attempted rates within the standard 45-min time limit and the 55-min extended time allocated for this study.
- The degree to which the effects of the DL test booklet accommodation are due to the additional time versus the availability of the items in both Spanish and in English through multivariate regression analyses.
- The equivalency of the test booklets with one another by comparing students’ mean levels of performance and through DIF analyses.
Timing Accommodation

Our study design had built in a within-group factor to evaluate the speededness of the test across the four different groups of students (as defined by crossing NAEP inclusion status with booklet type). However, the field supervisors noted that most students finished each of the two test sessions within the first 30 min allocated. The test data bore out these observations. Although the attempt rate was over 96% for each of the four study groups, the quality of the effort was not high: As the field supervisors observed, the students seemed to rush through and never looked back once they finished the last item. Only 17 of the 402 students used the time between the 45- and 55-min marks. Of these 17 students, 12 were in Group A (less than 3 years of instruction in English, administered the DL test booklet), 4 were in Group B (3 or more years instruction in English, administered the DL test booklet), and 1 was in Group C (3 or more years instruction in English, administered the English-only test booklet). Of the 12 Group-A students, 5 were sampled to be in the focus groups. These students’ use of the additional time might reflect additional motivation generated by the anticipation of the $25 reward they were to receive for participating in the focus group. The low-stakes nature of this assessment and its attendant lack of motivation prohibited us from conducting any further analyses of the timing accommodation.

DIF

As discussed in our section on recruitment, the timing of the test administration conflicted with schools’ busy calendars, resulting in a much smaller sample than we had hoped to obtain. Because the sample sizes per group were insufficient for conducting full-fledged DIF analyses, we do not report these results here. We did, however, use DIF and classical item analyses in a very circumscribed manner: to select items for the cognitive interview study (see Study 3).
Differences in mean raw scores. Our first step in examining the equivalence of the test booklets was to compare the four groups of students in performance. As shown in Tables 3 and 4, the native English speakers (Group D) had the highest mean raw score. The low mean performance levels shown in Table 3 are notable and are further testimony (along with the reports of the field supervisors) to the limited way in which these students cognitively engaged in this task. These low mean levels of achievement are comparable to the distribution of scores in NAEP eighth-grade mathematics administrations and illustrate the motivation problem associated with low-stakes testing. Table 4 is a more detailed display of the results of the group comparisons (mean differences, \( t \) values, and effect sizes) for the total, multiple-choice, and constructed-response scores. The results of this initial step in our analyses can be summarized as follows:

- Native English speakers (Group D), on average, outperformed native Spanish speakers (Groups A, B, and C).

### TABLE 3

<table>
<thead>
<tr>
<th>Student Groups Based on NAEP Inclusion Criteria</th>
<th>MC(^a) M</th>
<th>SD</th>
<th>CR(^b) M</th>
<th>SD</th>
<th>Total(^c) M</th>
<th>SD</th>
<th>Test Score Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A(^d)</td>
<td>17.17</td>
<td>6.47</td>
<td>10.09</td>
<td>5.42</td>
<td>27.26</td>
<td>11.05</td>
<td>.86</td>
</tr>
<tr>
<td>Native Spanish speakers, less than 3 years of academic instruction in English, received DL booklet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group B(^e)</td>
<td>18.19</td>
<td>6.95</td>
<td>12.46</td>
<td>5.76</td>
<td>30.65</td>
<td>11.74</td>
<td>.87</td>
</tr>
<tr>
<td>Native Spanish speakers, 3 years or more academic instruction in English, received DL booklet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group C(^f)</td>
<td>19.43</td>
<td>7.70</td>
<td>13.49</td>
<td>6.51</td>
<td>32.92</td>
<td>13.36</td>
<td>.90</td>
</tr>
<tr>
<td>Native Spanish speakers, 3 years or more academic instruction in English, received English-only booklet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group D(^g)</td>
<td>20.98</td>
<td>6.88</td>
<td>14.44</td>
<td>6.35</td>
<td>35.41</td>
<td>12.50</td>
<td>.90</td>
</tr>
<tr>
<td>Native English speakers, received English-only booklet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. NAEP = National Assessment of Educational Progress; MC = multiple-choice items; CR = constructed-response items; DL = dual-language.

\(^a\)Maximum possible score = 45. \(^b\)Maximum possible score = 32. \(^c\)Maximum possible score = 77. \(^d\)\( n = 127 \). \(^e\)\( n = 74 \). \(^f\)\( n = 119 \). \(^g\)\( n = 82 \).
<table>
<thead>
<tr>
<th>Comparison</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>M Difference</th>
<th>t Value</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total test performance</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td>127</td>
<td>27.26</td>
<td>11.05</td>
<td>–3.39</td>
<td>–2.05*</td>
<td>–.30</td>
</tr>
<tr>
<td>Group B</td>
<td>74</td>
<td>30.65</td>
<td>11.74</td>
<td>–5.66</td>
<td>–3.60**</td>
<td>–.46</td>
</tr>
<tr>
<td>Group C</td>
<td>127</td>
<td>32.92</td>
<td>13.36</td>
<td>–8.15</td>
<td>–4.94**</td>
<td>–.70</td>
</tr>
<tr>
<td>Group D</td>
<td>82</td>
<td>35.41</td>
<td>12.50</td>
<td>–10.70</td>
<td>–5.94**</td>
<td>–1.15</td>
</tr>
<tr>
<td>Group A</td>
<td>119</td>
<td>32.92</td>
<td>13.36</td>
<td>–2.27</td>
<td>–1.20</td>
<td>–.18</td>
</tr>
<tr>
<td>Group B</td>
<td>127</td>
<td>30.65</td>
<td>11.74</td>
<td>–4.77</td>
<td>–2.45*</td>
<td>–.39</td>
</tr>
<tr>
<td>Group C</td>
<td>82</td>
<td>35.41</td>
<td>12.50</td>
<td>–2.50</td>
<td>–1.34</td>
<td>–.19</td>
</tr>
<tr>
<td>Group D</td>
<td>82</td>
<td>35.41</td>
<td>12.50</td>
<td>–2.50</td>
<td>–1.34</td>
<td>–.19</td>
</tr>
<tr>
<td><strong>Performance, multiple-choice</strong>&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td>127</td>
<td>17.17</td>
<td>6.47</td>
<td>–1.02</td>
<td>–1.05</td>
<td>–.15</td>
</tr>
<tr>
<td>Group B</td>
<td>74</td>
<td>18.19</td>
<td>6.95</td>
<td>–2.26</td>
<td>–2.50*</td>
<td>–.32</td>
</tr>
<tr>
<td>Group C</td>
<td>127</td>
<td>19.43</td>
<td>7.70</td>
<td>–3.81</td>
<td>–4.05**</td>
<td>–.57</td>
</tr>
<tr>
<td>Group A</td>
<td>127</td>
<td>18.19</td>
<td>6.95</td>
<td>–4.34</td>
<td>–4.43**</td>
<td>–.67</td>
</tr>
<tr>
<td>Group B</td>
<td>74</td>
<td>19.43</td>
<td>7.70</td>
<td>–1.24</td>
<td>–1.13</td>
<td>–.17</td>
</tr>
<tr>
<td>Group C</td>
<td>82</td>
<td>18.19</td>
<td>6.95</td>
<td>–1.55</td>
<td>–1.46</td>
<td>–.21</td>
</tr>
<tr>
<td>Group D</td>
<td>82</td>
<td>19.43</td>
<td>7.70</td>
<td>–1.55</td>
<td>–1.46</td>
<td>–.21</td>
</tr>
<tr>
<td><strong>Performance, constructed-response</strong>&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td>127</td>
<td>10.09</td>
<td>5.42</td>
<td>–2.36</td>
<td>–2.92**</td>
<td>–.43</td>
</tr>
<tr>
<td>Group B</td>
<td>74</td>
<td>12.46</td>
<td>5.76</td>
<td>–3.39</td>
<td>–4.43**</td>
<td>–.57</td>
</tr>
<tr>
<td>Group C</td>
<td>127</td>
<td>13.49</td>
<td>6.51</td>
<td>–4.34</td>
<td>–5.29**</td>
<td>–.75</td>
</tr>
<tr>
<td>Group A</td>
<td>127</td>
<td>14.44</td>
<td>6.35</td>
<td>–1.03</td>
<td>–1.11</td>
<td>–.16</td>
</tr>
<tr>
<td>Group B</td>
<td>74</td>
<td>12.46</td>
<td>5.76</td>
<td>–1.98</td>
<td>–2.03*</td>
<td>–.33</td>
</tr>
<tr>
<td>Group C</td>
<td>119</td>
<td>13.49</td>
<td>6.51</td>
<td>–0.95</td>
<td>–1.03</td>
<td>–.15</td>
</tr>
<tr>
<td>Group D</td>
<td>82</td>
<td>14.44</td>
<td>6.35</td>
<td>–0.95</td>
<td>–1.03</td>
<td>–.15</td>
</tr>
</tbody>
</table>

<sup>a</sup>Maximum possible score = 77. <sup>b</sup>Maximum possible score = 45. <sup>c</sup>Maximum possible score = 32. 
*<i>p < .05</i>. **<i>p < .01</i>. ***<i>p < .001</i>. 
• Group A (native Spanish speakers, less than 3 years of academic instruction in English, using the DL booklet) scored significantly lower than the other three groups.

• Group B (native Spanish speakers, 3 or more years of academic instruction in English, using the DL booklet) and Group C (native Spanish speakers, 3 or more years of academic instruction in English, using the English-only booklet) were not significantly different from one another in mean test performance.

Because there were few statistically significant group differences and because as a whole, the effect sizes of the differences were small, we can infer that no systematic bias due to translation exists.

**Differences in adjusted mean scores.** The next step in our quantitative analyses was a multivariate examination of student performance in which we included language proficiency as a covariate. Our investigation of students’ responses to the items in the background questionnaire made us question the accuracy of students’ group assignments. For example, some students in Group A reported that they used only the English versions of the items in the DL test booklet. The correlations we computed between the language background data and test performance indicated that English-language proficiency was the factor most closely related to achievement, so we used this as a covariate in our analyses. Correlations between English-language proficiency and performance were .29, .23, and .31 for the total test performance, multiple-choice items, and constructed-response items, respectively (all significant at the conventional $p = .05$).

Table 5 displays the mean levels of self-reported proficiencies in English and in Spanish for each of the four student groups. Proficiency was calculated as the sum of students’ ratings of their abilities to read, write, listen, and speak in each of the two languages (Cronbach’s alphas across all groups were .93 for English proficiency and .94 for Spanish proficiency). The scores reported here range from 4 to 16 and are coded so that higher values represent higher proficiency ratings.

As can be seen in Table 5, the English proficiency ratings for the native Spanish speakers (Groups A, B, and C) are rather high. Indeed, the mean self-reported English proficiency ratings did not significantly differ between Groups B, C, and D. Because the average level of English proficiency was about the same between the groups, we questioned whether comparisons being made between the groups were meaningfully assessing the effect of English-language ability on the use of a DL test booklet. One way to isolate this effect is to consider the relation between a student’s English-language ability and performance and to see if that relation differed by type of booklet. Consequently, we conducted regression analyses to examine the effects of booklet type, language proficiency, and the interaction between booklet type and proficiency on test performance.
Our first series of regressions combined Groups A and B (students in the DL booklet condition) and contrasted them with the combined Groups C and D (students in the English-only booklet condition). We ran the same model (booklet type, English proficiency, and the Booklet × English Proficiency interaction term as predictors) three times to examine the effects of these predictors on (a) total test performance, (b) performance on multiple-choice items, and (c) performance on constructed-response items. These regression equations explained 10% of the variance in total test scores, 7% of the variance in multiple-choice scores, and 12% of the variance in constructed-response scores.

With respect to total test performance, we found no main effect for booklet ($\beta = -.42, \text{ns}$) but found a significant main effect for English proficiency ($\beta = .18, p = .006$). The booklet × English proficiency interaction was not statistically significant ($\beta = .57, p = .06$).

We also ran these regressions using Groups B and C only to try to tighten the focus on booklet type and English proficiency (these groups were the native Spanish speakers who had 3 or more years of academic instruction in English; Group B used the DL booklet and Group C used the English-only booklet). The patterns that emerged from these analyses showed that at every level of English proficiency, students who used the DL test booklet scored lower than those who used the Eng-
lish-only booklet. These data suggest that a DL test booklet may not be an aid and perhaps may even be a hindrance. Figure 1 displays the means, adjusted for English proficiency, for the two groups across the three performance measures.

Consistent with these findings, Abedi et al. (1997) found a similar type of interaction in their linguistic modification study. Abedi et al. (1997) found that LEP students administered the English or modified English test booklets scored higher than those administered the Spanish test booklet. This interaction held even after controlling for reading proficiency. Abedi et al. (1997) suggested that LEP students perform best on math tests when the language in which they are tested matches the language in which they are instructed. Abedi et al.’s (1997) follow-up analyses showed that students taught math in Spanish performed best when given the Spanish test booklet. This may be the case for the DL study as well. Although we did not have this information for all of our study participants, the majority of our focus group participants (see Study 2) reported that their math classes were conducted in English.

![Figure 1](attachment:image.png)

**FIGURE 1** Mean performance levels for Group B and Group C, adjusted for self-reported English proficiency.

*Note.* MC = multiple choice; CR = constructed response.
Our own follow-up analyses focused on the language in which students used to answer the test questions. We recognize that a self-reported language proficiency measure is very limited and is highly subject to biases such as social desirability, so we decided to add the language in which students answered the test as a covariate. That is, perhaps the seemingly inhibitory effect of the DL booklet on student performance was the result of a model that was improperly specified: That is, the model failed to address whether students actually took advantage of the accommodation that was made available. Our next set of analyses was conducted to address this issue.

**Group performance by language used to answer test questions.** We sought to refine further our understanding of the patterns we observed in the previous set of regressions by factoring in the language in which students answered the test questions. We considered this to be a more precise examination of the DL booklet’s effect on performance. The preceding analyses were overly simplified in that the comparisons were between students in the DL and the English-only test booklet conditions: Whether students made use of the accommodation by answering the Spanish version of the items was not considered. Given the high mean levels of English proficiency reported by our participants, the DL format may even have been irrelevant to some. Accordingly, language used to answer the test would serve as an indicator of the appropriateness of the DL test booklet as an accommodation. To investigate this issue, we redistributed the Group A and Group B students (those who were administered the DL test booklets) into

- Those who chose to answer in Spanish 90% or more of the items (Group AB–Spanish; \(n = 116\): 109 from Group A, 7 from Group B).
- Those who chose to answer in English 90% or more of the items (Group AB–English; \(n = 77\): 13 from Group A, 64 from Group B).

We then compared the performance of the Group AB–Spanish, Group AB–English, and Group C students with one another and found no significant differences in performance once English proficiency was controlled for. The results here and in the preceding analyses testify to the importance of English proficiency in students’ performance and indeed, of the complexity in determining for whom a testing accommodation is most appropriate. The lack of significant differences between native Spanish speakers’ (i.e., Groups A, B, and C) test performance also suggests that the two booklet types are comparable once English proficiency and language used to answer test questions are accounted for.

---

2Five students from Group A and three students from Group B were eliminated from this analysis because these students showed a mixed pattern of responses, answering a substantial proportion of items in English and a substantial proportion of items in Spanish.
Conclusions

Establishing equivalence between the two types of test booklets posed a challenge because of the sample size. The results of the regressions suggest that proficiency in English is an important factor in determining the influence of the DL test booklet accommodation on students’ mathematics test performance. Although there was a weak trend suggesting that the DL test booklet may have somewhat inhibited student performance, this pattern disappeared once we accounted for the language the student used to respond to the test items.

The results of these analyses illustrate the complexity of evaluating the effectiveness of an accommodation. We had initially cast the question in terms of “Which is better for LEP students, the DL or English-only test booklet format?” Closer inspection of (a) the interaction plots and (b) the analyses that included language used to answer test items suggest that the more appropriate question is, “For whom is a dual-language test booklet accommodation most effective?”, or “Who is most likely to use and benefit from a dual-language test booklet?” Furthermore, we found no differences in native Spanish speakers’ test performance once we had accounted for both English proficiency and language used to answer the test questions, which indicated comparability between the DL and English-only test booklets.

The results reported here suggest that implementing additional criteria on top of (and perhaps instead of) the 3 years of academic instruction in English rule merits consideration. The inclusion and accommodation policies of states and other large-scale assessments span a wide range of practices (e.g., language assessment scales, years in the United States, school performance, teacher observations/recommendations; NCES, 1997; Olson & Goldstein, 1997; Rivera et al., 2000) and go well beyond the 3 years of academic instruction in English criterion. Based on the results of our study and the complex picture that emerged, inclusion and accommodation policies must carefully weigh the advantages of the precision in classification offered by additional or alternative inclusion criteria with the burden those criteria would place on participating schools and districts.

STUDY 2: FOCUS GROUPS

The quantitative data gathered in Study 1 offer insights regarding students’ performance on the two types of test booklets. It is also important to conduct qualitative research to complement the performance data. Proponents of focus groups point to the rich data that are obtained, a direct result of the group dynamics unique to this methodology (e.g., Morgan, 1988). We consider focus groups also as a complement to the cognitive interviews reported under Study 3. In contrast to a cognitive interview think-aloud task, which can be challenging to some participants, focus groups provide a format that may be more comfortable for students: discussing
school experiences among their peers. The purpose of the focus groups was to garner valuable insights into the quality of the translation, content, format, and administration of the test as well as students’ use of the DL test booklet and its perceived utility.

Research Questions

It is important to note that the questions listed following were used as general guidelines for discussion rather than as a formal interview. Although a focus group is conducted as a group interview, the administrator’s role is to facilitate discussion and stimulate dialogue rather than to ask each participant to provide a response to every item in the protocol. The questions listed here and in the protocol were meant to generate comments and reactions and were not necessarily presented in the order listed:

- What were students’ perceptions of the quality of the translation used for the DL test booklet? Did they encounter any unfamiliar phrases or terms? What items were particularly challenging?
- How did students use the DL test booklet? How useful was the DL booklet?
- What difficulties did students have with the English-only test booklet? Did they encounter any unfamiliar phrases or terms? What items were particularly challenging?
- What were students’ opinions regarding alternative testing accommodations?
- How challenging was the test content? How much effort did the students make?
- How effective were the booklet instructions? What changes would they suggest?

Method

Participants. Native Spanish speakers (students from Groups A, B, and C) were eligible to participate in the focus groups. We randomly chose 6 of the 10 participating schools for the focus groups, recruiting one type of student group from each school (e.g., Group A focus groups were conducted in Schools 4 and 10; Group B focus groups were conducted in Schools 1 and 6). Students who were eligible to participate in focus groups were asked to bring home permission slips for their parents to sign, and only students who had signed permission slips were allowed to participate. In 1 of the 6 randomly selected schools, 5 of the Group A students sampled appeared for the focus group, but only 2 had signed consent forms. We did conduct the focus group with the 2 girls so as not to disappoint them. To compensate for this turn of events, we randomly selected another school and re-
crused additional Group A students. All told, our 68 focus group participants were distributed in the following manner:

- Group A (less than 3 years academic instruction in English, administered the DL booklet): $n = 31$.
- Group B (3 or more years academic instruction in English, administered the DL booklet): $n = 16$.
- Group C (3 or more years academic instruction in English, administered the English-only booklet): $n = 21$.

**Procedures.** The focus groups were conducted after the test, while the test-taking experience was still fresh. Students were given a 30-min break after the test administration, during which the field administrators prepared the room (e.g., set up the camera and chairs, laid out the refreshments). As students returned to the test administration room, permission slips were gathered and students were invited to partake of the snacks we had provided. After the focus group was completed, students were given $25 cash each and our thanks for their contributions to the study.

The focus groups were led by bilingual interviewers and were videotaped. To ensure accuracy, a professional translation company transcribed the videotapes. We should note that because we had to use the classrooms provided by the schools, there was often a great deal of background noise (e.g., students in the hallway, bells announcing the end or beginning of a class period, doors slamming). The videotape of one of the Group C focus groups ($n = 10$) could not be transcribed because of the poor quality of the audio. Accordingly, the results reported here are based on the six remaining focus groups. Because Study 2 was intended to be descriptive and because its focus is on the reactions of Groups A and B to the DL booklet, we remain confident that the loss of data from one of the Group C focus groups did not compromise the main findings of this study.

**Results and Discussion**

As part of the icebreaker, we elicited a discussion of students’ language backgrounds. Group A students (those with fewer than 3 years of academic instruction in English) generally spoke Spanish with their family and friends. Group B and C students were more heterogeneous, some reporting that they spoke Spanish only at home and English with friends, whereas others described using a mix of both Spanish and English with friends and family. When asked about what language was used in their math classes, the majority reported that their mathematics instruction was delivered in English. Several students stated that their grades in math were not very good. One student explicitly noted that it was nice that they got chosen to participate in this study instead of “just the smart kids.” We also inquired
about the level of effort students put into answering the test questions: All the focus
group participants admitted that they did not try as hard as they would have on a
“real” test.

**Language and content.** One of the main purposes of conducting focus
groups was to obtain feedback from students about the quality of the translation.
We were particularly interested several phrases and words from the test booklet be-
cause these were objects of some debate during the translation process. For exam-
ple, we were concerned that *desdoblado*—used in the phrase *papel desdoblado*
(unfolded paper)—might not be a familiar term.

We asked students how they interpreted these various phrases, and none had dif-
ficulty in defining the words and using the connotations we had intended. The dif-
ficulties in understanding expressed by the students appeared to have more to do
with familiarity with the mathematical concepts rather than the quality of the trans-
lation. Students from each of the three groups (Groups A, B, and C) described the
words *bisector, perpendicular,* and *parallelogram* (for which the Spanish versions
are virtually identical to the English words) as unfamiliar.

As a whole, no difficulties were reported with the Spanish that was used in the
translation. The use of the *usted* over the *tu* form of address was not of concern to
any of the Group A or Group B students. One Group B student joked about the is-
 sue by saying that she preferred *vos* (a form of address found in classic Castilian
and rarely used outside of Spain or Argentina). Students did note that it was helpful
to have several versions of the same word within an item to ensure that readers
from different countries would be able to understand the question (e.g., “*El costo
de arrendar (alquilar) una motocicleta …*”).

Students from all three groups found the content appropriate to what they were
learning in their math classes. Although terms such as *bisector, perpendicular,* and
*parallelogram* gave the students difficulty, the focus group participants all agreed
that the test was quite reasonable, if not downright easy. We pursued this paraodoxi-
cal feedback in a Group A and in a Group B focus group. Several of the students
explained that compared to the local district exam “where the question has all these
parts and you have to explain everything,” our test was much less demanding. An-
other explanation for the paradoxical comments was that the intent of the items
was clear; it was simply that the students could not remember the definitions of
words such as *perpendicular.* It appears that students begin from the premise
“math is hard” and expect the baseline level of difficulty to be high; instead, it is
their perceptions of the relative effort required in answering the test items that form
the evaluation of test difficulty. Therefore, although students reported that there
were several concepts with which they were unfamiliar, the test was still consid-
ered “easy” because it did not require them to perform many substeps or do a great
deal of writing. The items in Session 1 were considered to be easier than the items
in Session 2 (which called for the use of a calculator and a ruler). The two extended
constructed-response items (Session 1, Question 29, Carla and Maria’s subtraction
game, and Session 2, Question 29, the broadcast areas of two radio stations) were
consistently described as difficult.

Some of the questions in the background questionnaire were thought to be odd.
Several Group B students wondered why we wanted to know how many Eng-
lish-language books and magazines were in the home. A Group B student stated
that it was a good idea to have questions like this (e.g., perceptions of language flu-
ency, demographic information such as years in the United States). She astutely
noted that it would be important to know this sort of information to understand
why students answered incorrectly. The example she gave was that a student might
not know the answer because it was taught differently in one’s native country.

Format and administration. Because this test form and its accompanying
administrative procedures had never been fielded, we requested that students com-
ment on the test administration. When asked about the length and content of the
test booklet instructions, students agreed that these were fine. The instructions
were necessary features of the test booklet, and students reported that the text was
direct and to the point. Reactions to the calculator instructions were more mixed.
Some students thought that these were perfectly appropriate, but some Group B
students thought that calculator instructions were unnecessary and should be de-
leted from the protocol.

Students did not have much feedback to offer with regard to format. The test we
administered was similar to others they had taken, and so the test was simply par
for the course. One Group A student was of the opinion that the booklet layout was
too formal and plain and could benefit from the addition of graphics on the cover.

Uses and utility of the DL booklet. The DL booklet was very well received.
Although students tended to use the questions and to answer in one language, they
found it helpful to have the second language on the other page to use as a compre-
hension check. One student in Group B noted that this format was particularly use-
ful because one never knew which word one might not understand. That is, if one
were doing the test in English, one could use the Spanish to check one’s under-
standing. By the same token, if one were doing the test in Spanish and encountered
a problematic word, one could then use the English to clarify the meaning. Another
Group B student added to this observation by noting that the DL format would be
helpful for students who are not proficient in reading or writing in Spanish (despite
being fluent Spanish speakers). She noted that if a student could not read the big
words in Spanish, the student could use a combination of the English and Spanish
during a test.

The DL format was highly endorsed: Students believed that they were better
able to demonstrate their understanding of the test items by having the questions
available in the two languages. Several students were of the opinion that high-stakes tests such as the Florida Comprehensive Assessment Test and the Scholastic Assessment Test should offer DL booklets. Others recommended that we develop DL tests in all subject areas and expand our scope to other languages such as Portuguese. We did show Group C students a copy of the DL test booklet, which was met with interest. One Group C student said that she was not sure if having the items in Spanish would have helped, but she was the sole dissenter. Another Group C student noted that he could not determine whether his difficulty with mathematics was due to the language or to the subject.

Group A and Group B students (who used the DL booklet) preferred the DL format to a Spanish-only version. They explained that a Spanish-only version might disadvantage students who could speak but not read or write in Spanish and that having items in both Spanish and English can help one learn English. Students strongly preferred the DL format to having a Spanish–English dictionary and an English-only test booklet. They considered having to look up words in the dictionary to be a tremendous disadvantage during a timed test. They also observed that having a Spanish–English dictionary available would not be helpful to students who can speak but are not literate in Spanish. Another student remarked that dictionary definitions do not always clarify the meaning of a word.

Conclusions

The group discussion format of focus groups was successful in eliciting students’ opinions and observations about the test content and the DL booklet accommodation. Although some students were not as vocal as others, the videotapes show that the less talkative students were still engaged in the discussion, following what was said and expressing agreement or disagreement with nods or shakes of their heads.

The DL test booklet was viewed positively. Although students administered the DL booklet tended to use one language predominantly, all considered the availability of the second language a benefit. Indeed, we were encouraged to expand DL testing into the other content areas as well as to other languages. These findings are an interesting contrast to the quantitative analyses, which indicated that the DL test booklet has a slight negative effect on the performance of students who have higher levels of English proficiency. However, it may be that the Group A and B students in our focus groups were those at the lower levels of English proficiency.

The qualitative data collected in this study also indicate that the translation of items from English to Spanish was effective. Phrases and words that we anticipated might be ambiguous or unfamiliar posed no difficulties for the students. Vocabulary was troublesome only to the extent that students were unfamiliar with the math concepts involved (e.g., bisector). We view the results of these focus groups as additional evidence for a successful translation.
STUDY 3: COGNITIVE INTERVIEW STUDY

This subtask addresses the need to gather linguistic and psychological evidence to testify to the accuracy of a test adaptation (Hambleton, 1994; International Test Commission, 1994; van der Vijver & Hambleton, 1996). Translation issues, such as using proper grammar and vocabulary and avoiding cultural bias, were explored by examining a sample of the 60 test items from the perspective of their potential users. We examined closely how students interpreted the items and what strategies they used to solve the items. This information was then used to determine whether the items were translated properly and whether they measured the constructs intended by the item writers. The cognitive interview study was conducted as a follow-up to Study 1, in which we analyzed test data to evaluate the comparability of the DL and English-only test booklets, and as a complement to Study 2, in which we conducted focus groups to obtain information on students’ reactions to the test and the test format.

Method

Participants. A total of 18 eighth-grade students participated in the study. Among the participants, 5 were native Spanish speakers with less than 3 years of academic instruction primarily in English (usually with fewer than 3 years in the United States; Group A), and 3 were native English speakers (Group D). Ten students were native Spanish speakers with 3 or more years of academic instruction in English (most since first grade); half were administered the DL test booklet (Group B) and half were administered the English-only test booklet (Group C). All native Spanish speakers were of Mexican descent. Two of the native English speakers were White, and the third was East Indian. There were 10 girls and 8 boys in the sample.

To recruit participants, schools and community centers in northern California were visited and flyers were distributed during August and September of 2000. All students who expressed interest in participating were included in the study only after parental consent was obtained. The comparisons of performance across the four groups were expected to provide information to help understand DIF results and to illustrate the use, benefits, and drawbacks of the DL test booklet for students who are native Spanish speakers.

Selection of items. Of the items used in Study 1, 14 were selected for the cognitive laboratories. The first criterion was related to DIF indicators. We used the Mantel–Haenszel method for DIF, and the matching was done by dividing the sample into five groups based on students’ total score (0%–20%, 21%–40%, 41%–60%, 61%–80%, 81%–100%).
41%–60%, 61%–80%, 81%–100%). Although we acknowledge that this combination of “thick matching” of ability levels and large group mean effect size differences sets the conditions for overidentification of DIF, we cast a wide net for selection of items to include in the cognitive interview subtask. We required also that an item had to demonstrate one of the following three subcriteria:

- Significant DIF (at the \( p = .05 \) level) for the Group A (native Spanish speakers, less than 3 years of academic instruction in English, DL booklet) versus Group D (native English speakers, English-only booklet) comparison;
- Significant DIF for the Group B (native Spanish speakers, 3 or more years of academic instruction in English, DL booklet) versus Group C (native Spanish speakers, 3 or more years of academic instruction in English, English-only booklet) comparison; or
- Very poor discrimination for at least one of the four groups in which very poor was defined as having a low (less than .05) or negative biserial correlation with the total score, using classical item analysis statistics.

Altogether, 16 items met at least one of these three subcriteria. In addition, Item 29 from Session 1 was also included in the initial list of items recommended for the cognitive laboratories, despite the fact that it did not technically meet any of the

---

3The classifications for flagging items for DIF are based on the following rules. For multiple-choice items, the item classifications were based on the Mantel–Haenszel (MH) chi-square and the MH delta (\( \Delta \)) values:

- An item was classified into the “C” category if the absolute value of the MH delta value (i.e., \(|\Delta|\)) was significantly greater than 1 and also greater than or equal to 1.5.
- An item was classified into the “B” category if the MH delta value (\( \Delta \)) was significantly different from 0, and either the absolute value of the MH delta (\(|\Delta|\)) was less than 1.5 or the absolute value of the MH delta (\(|\Delta|\)) was not significantly different from 1.
- An item was classified into the “A” category if the delta value (\( \Delta \)) was not significantly different from 0, or the absolute value of delta (\(|\Delta|\)) was less than or equal to 1.

For constructed-response items, the item classifications were based on the MH chi-square and the SMD index (standardized mean difference, a summary statistic that compares the means of the reference and focal groups, adjusted for differences in the distribution of reference and focal group members across the values of the stratification variable):

- An item was classified into the “CC” category if the MH chi-square \( p \) value was less than .05, and the absolute value of SMD divided by standard deviation of the item score (i.e., \(|SMD/SD|\)) was larger than .25.
- An item was classified into the “BB” category if the MH chi-square \( p \) value was less than .05, and the absolute value of SMD divided by standard deviation of the item score (i.e., \(|SMD/SD|\)) was .17 to .25.

All other items were classified into the “A” category.
three preceding subcriteria. This item was included because it was one of the two extended constructed-response items on the assembled test, and it had a significance level of .052 for the Group B versus Group C comparison. All 16 of these NAEP items had nonsecure, public release status. The items that we identified as demonstrating DIF did not reveal any consistent patterns with regard to item type, question length, content strand, or ability tested.

The second criterion was related to other classical test statistics. To be included in the set of items recommended for use in the cognitive laboratories, the item had to have acceptable classical item statistics in Group D (native English speakers). We eliminated items that were too easy (percentage answering item correctly was greater than 95), were too hard (percentage answering item correctly was less than 20), had poor item–total biserial correlations for the correct response (biserial less than .15), or had item–total biserial correlations for one or more incorrect response options equal to .10 or greater.

Accordingly, 10 of the 16 items were chosen for use in the cognitive laboratories. Of these items, 6 are multiple choice, 3 are short constructed response, and 1 is extended constructed response. We decided to include also 4 control items in the cognitive interview protocol. Thus, the cognitive interview protocol consisted of 14 NAEP items. All 14 items chosen for use in the cognitive interview study had public release status; no secure NAEP items were used in this subtask.

Materials. Using these 14 items, we developed two types of student booklets: a DL booklet formatted in the same manner as described previously (i.e., the English version on the right side, the Spanish version on the left side) and an English-only booklet. The DL booklet included two introductory paragraphs in Spanish that students read at the beginning of the section. The interviewers used this to assess students’ Spanish reading proficiency. An interview protocol was developed to provide the interviewers with a semistructured script for the interview.

Procedures. Prior to interviewing students, interviewers participated in a 2-day training session in which they learned and practiced interviewing techniques for the cognitive interviews. Interviewers practiced the probing techniques used to prompt a student to think aloud without influencing students’ responses. Interviewers also received training on procedures for reporting data from the interviews (e.g., students’ Spanish reading proficiency, mastery of item content).

During the cognitive interview, the student chose the language in which the interview was conducted. The interview was videotaped and held over a 2-hr period that included a 15-min break for the student. First the interviewer explained the purpose of the study and what the student would be doing in the interview. The student was then trained in the process of thinking aloud using exercises created for this purpose. The interviewer evaluated each native Spanish-speaking student’s ability to read in Spanish by recording the number of words the student mispro-
nounced as he or she read aloud the two paragraphs in the test booklet that explained the purpose and components of the cognitive interview. We emphasized to the students that the session was not a test and that we were interested only in learning how students solve math problems.

The cognitive interview consisted of three phases:

- **Phase 1: Think alouds.** The student answered each item, and the interviewer used only general prompts to motivate the student to think aloud (e.g., “Remember to think aloud while answering the question,” or “You are doing a great job thinking aloud.”). The interviewer recorded the steps the student took in solving the item and any thoughts the student verbalized about what the item was asking.

- **Phase 2: Probing.** Interviewers explored the reasons behind a student’s response by asking the student item-specific questions. The probes enabled the interviewer to clarify exactly what solution path the student used and whether the student encountered any problems in answering the item.

- **Phase 3: Analysis of student performance.** Interviewers gave detailed reports of a student’s performance with the items on data recording forms.

**Research Questions and Data Analyses**

This study addressed the following research questions:

**Research Question 1.** Is there any evidence that the grammar or vocabulary of the Spanish-language items in the DL booklet was problematic (i.e., unclear or difficult to understand for the intended population)?

**Research Question 2.** Are there differences between the English and Spanish versions of the same item with respect to their content?

**Research Question 3.** Is there evidence that the groups employed different solution pathways when responding to the items? Are these differences in solution pathways due to differences created by translating the item into Spanish?

**Research Question 4.** Did native Spanish speakers with less than 3 years of academic instruction in English use the DL booklet differently than native Spanish speakers with 3 or more years of academic instruction in English?

To address the first research question, detailed descriptions of student responses for each item were analyzed to identify whether there were words or phrases that students had difficulty interpreting or explaining. With regard to the second research question, we identified discrepancies between student performance and
mastery of the construct(s) being assessed and examined the extent to which these discrepancies were attributable to misinterpretation of the item.

To address the third research question, we identified typical solution pathways for each group and compared how they differed across groups. We were particularly interested in seeing whether there were differences in the proportions of group members who relied on either algorithms (procedures with specific rules that guarantee a solution) or heuristics (general rules of thumb that do not guarantee a solution).

The final research question addresses the need to provide this type of accommodation (DL test booklet) for native Spanish speakers with 3 or more years of academic instruction in English. The extent to which these students used the Spanish items in the DL booklet was interpreted as evidence that these students found this type of accommodation helpful.

Results

**Student performance and mastery of content.** Overall summaries of student performance and mastery are presented in Table 6. The “Responses That Showed Mastery” column was included to document our interviewers’ judgments of whether a student grasped the mathematical content or concept. This grasp, however, may not have led necessarily to a correct response (e.g., the number of responses that showed mastery was greater than the number of correct responses for Groups B, C, and D). Distributions across groups of student responses suggest that

<table>
<thead>
<tr>
<th>Group</th>
<th>Correct Responses</th>
<th>Partially Correct Responses</th>
<th>Incorrect Responses</th>
<th>Total Responses</th>
<th>Responses That Showed Mastery</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>%</td>
<td>No.</td>
<td>%</td>
<td>No.</td>
</tr>
<tr>
<td>A(^a)</td>
<td>27</td>
<td>40</td>
<td>9</td>
<td>13</td>
<td>31</td>
</tr>
<tr>
<td>B(^b)</td>
<td>44</td>
<td>63</td>
<td>6</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>C(^c)</td>
<td>49</td>
<td>70</td>
<td>5</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>D(^d)</td>
<td>37</td>
<td>88</td>
<td>3</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

**Note.** The total responses column equals the number of students in the cell times the number of items (14, except for Group A, which had missing data on two items). The \( n \) of 67 = (5 students \( \times \) 12 items) + (4 students \( \times \) 1 item) + (3 students \( \times \) 1 item).

\(^a\) Native Spanish speakers with less than 3 years of academic instruction in English using a dual-language booklet; \( n = 5 \).

\(^b\) Native Spanish speakers with more than 3 years of academic instruction in English using a dual-language booklet; \( n = 5 \).

\(^c\) Native Spanish speakers with more than 3 years of academic instruction in English using English-only booklet; \( n = 5 \).

\(^d\) Native English speakers using English-only booklet; \( n = 3 \).
groups were not equivalent with respect to their math performance. Group D (native English speakers using an English-only booklet) showed the highest level of math achievement, whereas Group A (native Spanish speakers with less than 3 years of academic instruction in English) showed the lowest level.

**Response time.** We recorded the amount of time that students took to respond to the items in Phase 1 of the interview, in which the student responds to the item without any probing by the interviewer. Students in Group D (native English speakers, English-only booklet) took the least time (an average of 70.1 sec for each item, $SD = 58.0$), whereas students in Group A (native Spanish speakers, less than 3 years instruction primarily in English, DL booklet) took the most (an average of 137.4 sec, $SD = 142.4$). Group B (native Spanish speakers, 3 or more years instruction primarily in English, DL booklet) and Group C (native Spanish speakers, 3 or more years instruction primarily in English, English-only booklet) took similar amounts of time to answer each item (Group B $M = 103.7$ sec, $SD = 78.9$; Group C $M = 109.3$ sec, $SD = 105.8$). The large standard deviations in response time for Groups A and C are noteworthy and are due to students’ use of heuristics rather than algorithms. Certain students did not know the appropriate mathematical concepts and operations called for by the item. Instead, they relied on a time-consuming, trial-and-error approach to determine a correct answer from the given response options.

**Quality of the translation.** With regard to the first research question, we found no evidence that the grammar or vocabulary of the Spanish items was technically incorrect. However, there were two phrases in the Spanish version of one item that seemed unclear, particularly for students with fewer than 3 years of academic instruction in English. These phrases were *voltear el triángulo de arriba sobre la línea l* (flipping the above triangle over the line l) and *cuál de las siguientes figuras* (which of the following figures).

We also observed that the wording and language structure for another item was confusing for native Spanish speakers in both the English and Spanish versions of the item. The multiple use of the passive voice in a conditional clause could have made the item difficult to comprehend: “¿Cuánto aumentaría 217 si el dígito 1 se reemplaza por el dígito 5?” (“By how much would 217 be increased if the digit 1 were replaced by a digit 5?”).

For a third item, we found that the students either ignored or forgot to comply with the instruction to round their answer to the nearest centimeter. Although the phrase *al centímetro más próximo* is a proper translation of the phrase *to the nearest centimeter*, the language seemed somewhat foreign to the students. The native Spanish speakers seemed more familiar with the phrase *round to the centimeter* or *redondear al centímetro*. 
Understanding of item content. The only item that suggested misunderstanding of the item’s content was the “flipping the above triangle over the line l” question. Group A students (those with less than 3 years of academic instruction in English) seemed to interpret the item differently than other native Spanish speakers. This was the only item in which a discrepancy between a student’s mastery and performance can be explained by a misunderstanding of the item’s content. That is, despite the fact that students’ think-aloud responses demonstrated that they understood what the question was asking, some students still chose the incorrect response because they understood the item differently from the way the item writer intended.

Differences in solution paths. Students from different groups employed different solution paths when responding to the items. This was most noticeable in comparing solution paths for native Spanish speakers with less than 3 years of academic instruction in English (Group A) with native English speakers (Group D). Rather than linguistic issues, differences in the typical strategies used by these students appeared to be more closely related to the groups’ familiarity with the appropriate algorithms to use to arrive at the correct answers. This was the case for 6 of the 14 items.

Group A students tended to correctly solve the problems primarily through the use of heuristics, whereas Group D students solved these problems mostly through the use of appropriate algorithms. For three of the items, the heuristics used were unrelated to the theoretical constructs that the items purported to measure (i.e., the heuristics were general test strategies that allowed students to guess the right answer despite their lack of skills or knowledge). No consistent pattern was demonstrated in the content strands and abilities tested that would suggest why these groups would differ in their solution paths other than familiarity with mathematical algorithms.

Use of the DL test booklet by native Spanish speakers. Our final research question focused on how Groups A and B used the DL test booklet. We found that Group B students (native Spanish speakers with more than 3 years of academic instruction in English) used the English version of the items the majority of the time. Group A students (native Spanish speakers with less than 3 years of academic instruction in English) used the Spanish version all of the time. In 12 out of 70 instances, native Spanish speakers with more than 3 years of instruction primarily in English read the Spanish version of the item to check their understanding of the English item or to look for the meaning of specific words and phrases (e.g., bisector). Facility in reading Spanish did not seem to be the reason for reliance on the English versions of the items: These students demonstrated no problems in reading Spanish during the proficiency assessment given prior to the interview (i.e., no mistakes in pronunciation were reported for this group).
Discussion

This cognitive interview study collected in-depth qualitative information to help explain DIF for items that were administered in a DL or an English-only test booklet to diverse samples of eighth-grade students. The rich qualitative data obtained from these cognitive interviews suggest that differences in the psychometric properties of some target items may have been due to linguistic factors. The cognitive interview results reinforce the argument that evaluation of translated items should transcend evaluation of the external structure of an item (e.g., vocabulary, grammar). For example, differences in how students interpreted items M016301 (flipping triangle over line l) and M022801 (length of one of the longer sides of the rectangle) arose not because the translation was technically incorrect. In item M016301, the wording was not specific enough (the verb voltear seemed to introduce some ambiguity in meaning). In item M022801, translators failed to use a term that students were familiar with (the verb redondear [to round] instead of the phrase el centímetro más próximo [the nearest centimeter]). We also found problems with the use of cuál de los siguientes (which of the following) and aproximar (to approximate). These results are consistent with those of the 1994 Puerto Rico Assessment of Educational Progress, Technical Report (Anderson & Olson, 1996).

Apart from linguistic issues, we explored reasons why target items may have shown different psychometric properties when administered to groups that differed with respect to native language and levels of mathematical achievement. The results of the DL cognitive interview study indicated that native English speakers, who showed higher levels of mathematical achievement compared with the other groups, tended to employ the appropriate algorithms in their solution paths. Correct answers were attributable to mastery of the construct.

Conversely, native Spanish speakers, particularly those with less than 3 years of academic instruction in English, showed a lack of full mastery, often failing to use the appropriate algorithms. However, these students were sometimes able to select the correct responses through the use of heuristics. The additional time that native Spanish speakers in this study took to answer each item represented their trial-and-error approach to responding. An implication of these results is that perhaps we should consider a greater emphasis on using algorithms when teaching Spanish-speaking LEP students, particularly with regard to fundamental mathematics skills.

The final issue explored in this study was the use of the DL test booklet by native Spanish speakers with more than 3 years of academic instruction in English. We found that students recruited into Groups B and C were a heterogeneous group of individuals ranging from students who had taken courses primarily in English since preschool to those who had received instruction primarily in English only since the sixth grade. The students in this sample used the Spanish items in the DL
test booklet largely as a way to check their understanding of an item, whereas native Spanish speakers with less than 3 years of instruction primarily in English used the Spanish items all the time.

DISCUSSION

Summary of Results

Test adaptation/translation. Our translation procedures were guided by the recommendations and review of the test adaptation literature. Evidence for the quality of the translation was obtained through quantitative analyses, focus groups, and cognitive interviews. Quantitative analyses identified 10 of the 60 items as exhibiting DIF, but given the characteristics of our sample and the procedures that had to be used, overidentification was likely. These 10 items (along with four controls) were closely examined in the cognitive interview subtask, and although several phrases appeared to be ambiguous, there was no direct evidence of incorrect or inappropriate translation. Focus-group data were consistent with these findings. Students had no difficulty with phrases that had been the subject of some debate by the translation team. Indeed, students’ difficulties with the content or language appeared largely due to a lack of mastery of a concept (e.g., bisector, perpendicular) than a flawed translation.

The translation process involved careful comparisons of the original, forward-translated, and back-translated documents by a bilingual team representing diverse Latino backgrounds and possessing content area expertise. These procedures are similar to what is done in the Third International Math and Science Study, a large-scale assessment that requires translation of the same instrument into many languages. Although a larger sample size would have been optimal and would have allowed us to conduct more exhaustive and detailed quantitative analyses of the test items, the indications from the data available here are encouraging, and the translation appears to have been effectively and accurately done.

Effectiveness and appropriateness of the DL format as a testing accommodation. Our group comparison analyses were directed toward establishing the effectiveness of the DL booklet as a test accommodation and to evaluating the comparability of the two types of test booklets. As described previously, initial analyses suggested that after controlling for English proficiency, the DL test booklet somewhat hindered test performance. However, the follow-up analyses were conducted that accounted for both English proficiency and use of the test booklet (i.e., language in which students answered the test) helped to establish the compa-
rability of the two test booklets. That is, after controlling for self-reported English proficiency, we found no significant differences between

- Native Spanish speakers who answered the DL test booklet in Spanish.
- Native Spanish speakers who answered the DL test booklet in English.
- Native Spanish speakers who were given the English-only test booklet.

Although we initially framed our inquiry as “Which is better for LEP students, the dual-language or English-only booklet?,” our results suggest that the more refined question to pose is “For whom is a dual-language test booklet accommodation most effective?” A simple comparison of the two test booklet conditions appears to be too coarse: We must also consider whether students actually avail themselves of the accommodation (i.e., answered in Spanish if administered the DL test booklet). This raises the issue regarding appropriateness of accommodations and the challenge of accurately identifying students who should receive language accommodations. The basic rule, “3 or more years of academic instruction in English” may not be the most precise inclusion criterion. Our data indicate that greater accuracy in targeting students in need of language accommodations may be achieved by also including measures of English-language proficiency (as many states already do).

Equity is always an issue, and a concern some might have is of the DL format possibly according an unfair advantage to native Spanish speakers. On the basis of our results, the DL test booklet does not appear to offer any undue advantage. Instead, this accommodation seems to equalize performance by allowing those who need the accommodation to demonstrate their mathematical understanding to the best of their ability.

**Use and utility of the DL test booklet.** Although the test administration data, focus group, and cognitive interview results showed that students tended to use one language predominantly, the DL test booklet was considered to be a valuable and beneficial format. Of the 181 students offered this accommodation and who answered the question in the language background questionnaire at the end of the test administration, 155 (85%) considered the DL booklet to be useful or very useful. Comments from the focus group discussions were unanimously positive. Students reported that they felt better able to express their mathematical understanding when using a DL booklet, and we were urged to develop similar tests for the other subject areas and to provide translations in other languages. Students preferred this format to a Spanish-only booklet or an English-only booklet accompanied by a bilingual dictionary. The DL format was viewed as more equitable to LEP students who spoke but perhaps were not literate in Spanish and was considered to be a good way to help a student learn English.
Review of our interaction plots of the regression analyses showed that once we accounted for English proficiency and language used to answer the test questions (i.e., whether students actually made use of the accommodation), we found no differences in performance between native Spanish speakers (Groups A, B, and C). These results suggest not only that the DL and English-only test booklets are comparable but also that we need to think carefully about how best to identify which students would most benefit from this type of accommodation.

Given (a) the positive way that this testing accommodation was greeted by its target examinees and (b) the study findings that suggest comparability between the DL and English-only test booklets, a DL test booklet appears to be a promising and effective method for inclusion of Spanish-speaking LEP students in large-scale assessments.

**CONCLUSIONS**

The data reported here are limited in two respects. First, the modest sample size prohibited more statistically powerful modeling and analyses of the test data. Second, the low motivation of the examinees precluded consideration of the extended time accommodation and likely introduced noise that attenuated the statistical relations we observed.

Although we acknowledge the limitations of this study, we would like to offer two suggestions to those considering a DL test format as an accommodation for English-language learners. First, we advocate including bilingual mathematics teachers as members of the translation team. Because of their experience and familiarity with the target population, teachers can offer valuable advice regarding the appropriateness of the translation and recommend phrases to use for optimal comprehension (see also Hambleton and Patsula’s 1998 and 1999 works, which offer excellent guidelines for selecting translators). Second, we encourage the continued use of techniques such as cognitive laboratories and focus groups to obtain insights into how students interpret items. The best translation teams may be able to avoid problematic adaptations of items, and the most sophisticated quantitative analyses may be able to identify DIF. However, to uncover the reasons behind such difficulties requires input from the examinees themselves, which is best gained through the more qualitative methodologies.

Although the results of this study are not definitive, we believe that the data and procedures reported here are informative to others concerned about providing testing accommodations to English-language learners. We look forward to additional research in this important area in large-scale assessments.
ACKNOWLEDGMENTS

An earlier version of this article was presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana (April 2002). The full-length final report (200+ pages) on which this manuscript was based is available on request from Teresa Duncan.

REFERENCES


Copyright of Applied Measurement in Education is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.