DOCUMENT RESUME

ED 464 111 TM 033 792

AUTHOR	Duncan, Teresa Garcia; Parent, Lourdes del Rio; Chen, Lee; Ferrara, Steve; Johnson, Eugene
TITLE	Study of a Dual Language Test Booklet in 8th Grade Mathematics.
PUB DATE	2002-04-00
NOTE	44p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	English; Grade 8; *Junior High School Students; Junior High Schools; Language Minorities; *Mathematics; Spanish; *Test Construction; *Testing Accommodations
IDENTIFIERS	*Dual Language Text

ABSTRACT

This study addressed the effectiveness of a Spanish-English dual language (DL) test booklet in eighth-grade mathematics. The inclusion criteria of the National Assessment of Educational Progress were used to identify students who would qualify for an accommodation like the DL booklet. Incentives (periodicals, software, and study data) were used to secure the participation of 10 schools and a sample of 402 eighth graders. Results of the quantitative analyses illustrate the complexity of evaluating the effectiveness of an accommodation. The complex study suggested that 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 schools and districts. Quantitative analysis indicated that the dual language booklet has a slight negative effect on students with higher levels of English proficiency. Overall, 68 students participated in focus groups, and these groups showed that students viewed the DL booklet favorably. All students considered the availability of the second language a benefit. Eighteen students participated in the cognitive interviews, and their responses resulted in some suggestions for the preparation of translated test booklets. It is suggested that bilingual mathematics teachers be included in the translation team and that qualitative methods be used to determine how students respond to translated tests. Study results should be useful to those planning to provide testing accommodations to English language learners. (Contains 8 tables and 12 references.) (SLD)

Study of a Dual Language Test Booklet in 8th Grade Mathematics

Teresa García Duncan

Lourdes del Río Parent

Lee Chen

Steve Ferrara

Eugene Johnson

American Institutes for Research

Presented at a paper session, *Validity of Assessments for Linguistic Minorities and Students with Disabilities*, at the annual meeting of the American Educational Research Association, New Orleans, LA (April, 2002).

Please direct all correspondence to:

Teresa García Duncan
American Institutes for Research
1000 Thomas Jefferson St. NW
Suite 400
Washington, DC 20007
Email: tduncan@air.org

Study of a Dual Language Test Booklet in 8th Grade Mathematics

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 is to address the effectiveness of a Spanish-English dual language (DL) test booklet in eighth grade mathematics. This accommodation involves a booklet format where 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 quantitative analyses of students' performance, focus groups, and qualitative cognitive interviews (think-alouds) to determine whether there existed psychometric equivalence between the DL and English-only versions of this mathematics test. This study bears directly on the validity and value themes of this year's conference, as it addresses the accuracy and utility of the dual language format as a testing accommodation.

This study was conducted to assist the National Assessment Governing Board (NAGB) 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 dual language test booklet, and this study was directed toward answering a series of questions regarding the equivalency, quality, and usability of a dual language 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 dual language test booklet study:

- Study 1. Evaluation of the psychometric equivalence of the dual language and English-only booklets via traditional quantitative analyses (e.g., differential item functioning and 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 qualitative information on the validity of the translation and about how students used the dual language 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 dual language 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:

- Academic: Are the grammar and language structure used in the native language version correct?
- Cognitive: Do students understand the native language version of the test questions as a vehicle for assessing mathematics?
- Content: Is the content of the native language version of the test questions the same as the English version?
- Cultural: Is the native language version clear and acceptable to the various communities in the United States for whom this is the native language?
- Psychometric Equivalence: Is there a psychometric equivalence between the dual language version and the English-only version of the test?

Method

Study Design

Student Groups and Test Booklet Conditions. 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 dual language test booklet (Anderson, Jenkins & Miller, 1996). Indeed, similar criteria exist for other large-scale assessments, such as those administered by states (NCES, 1997; Olson & Goldstein, 1997; Rivera, Stansfield, Scialdone & Sharkey, 2000). Currently, students who have had fewer than three years of academic instruction in English are excluded from NAEP administrations, unless one of the following two conditions are met:

- The NAEP test is available in the student's native language (presently, Spanish is the only language other than English in which the NAEP is administered)

OR

- 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 comprised of native Spanish-speaking students who had fewer than three years of academic instruction in English, 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 dual language test booklet. However, in order to evaluate fully the effectiveness and utility of a dual language 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 individuals

who could use either a dual language or an English-only test booklet into the two booklet conditions and compare their performances. Native English speakers were also included in the study design to determine the adequacy of the time allocated for each session (the timing issue is addressed below).

Extended Time. Because items translated into Spanish typically increase the length of a question (i.e., more syllables and longer words), we were concerned about the speededness of the dual language test for the native Spanish speakers.[1] Speededness was also a potential issue with native English speakers, because the test administered here had never been fielded. As a result, an additional 10 minutes was built into the study design. The selection of 10 minutes as the additional time was largely based on logistical considerations regarding burden on the students and the schools. The test specifications stipulated that the test be administered in two 45-minute sessions. The additional 10 minutes for each session meant that the test administration, once time for instructions and completing the language background questionnaire were factored in, would entail more than 2.5 hours from each participant. Offering more than 10 minutes of additional time per session seemed to be unduly burdensome. In light of this, we determined that the data on "not reached" and "attempted" items available from the 55 minutes per session would suffice for our purposes.

Test booklet instructions informed students that they would have 55 minutes to complete each section. Following the procedures used in the Program for International Student Assessment (PISA; Organization for Economic Cooperation and Development (OECD), 1999) students began each of the two test sessions with a standard #2 pencil. At the 45-minute mark, students were asked to switch to a colored pencil. This within-subjects feature of the study design was used to enable us to examine two issues:

- The degree of speededness of both the English-only and dual language test booklets.
- The degree to which the effects of the dual language test booklet accommodation for LEP students may be attributed to the provision of extra testing time versus the provision of test items in both languages.

In sum, our study design involves two between-group factors (accomodation status and booklet condition) and one within-group factor (standard versus extended time). The design is displayed in Table 1. This balanced, mixed-model design allows us to address specifically (a) student performance across booklet conditions and (b) the adequacy and appropriateness of the allocated testing time. The results of our quantitative analyses are reported under Study 1.

Insert Table 1 about here

---

[1] Tests that are speeded contain more items than can be completed comfortably by the majority of examinees that would take the test under operational test administration conditions.

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 use simply as many schools as were willing to participate in the study. There was a tremendous amount of testing occurring in spring 2000 (particularly in 8th grade), with States and testing companies already having claimed large portions of schools' time for testing and other non-instructional activities. Despite districts' interest in participating in the dual language study, many schools had to deny our request, due to a sheer lack of time. The Council of the Great City Schools (CGCS) led our recruitment efforts. Given the relevance of dual language 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,[2] 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, as described in Table 1) in every school. However, not all students sampled appeared for the test administration. Participation rates within each of the four study groups are shown in Table 2. 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 dual language test booklet condition is exactly the same of students assigned into the English-only test booklet condition (69.79%).

Insert Table 2 about here

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

[2] 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", 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.

questions were chosen consisted of the 1990, 1992, and 1996 NAEP 8th grade mathematics item banks. Two experienced AIR 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.[3] To attain the desired test information function[4] (TIF), an iterative procedure was initiated: using the item statistics provided by NCES (items' p-values and IRT parameters) items were removed and replaced and the TIF re-calculated 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, four 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 dual language 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 comprised of six AIR 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), with respect to meaning, word and sentence complexity, and word and sentence length.

[3] For more detailed descriptions of VNT-M test specifications, please refer to *Voluntary National Test in Grade 8 Mathematics Test Specifications*, American Institutes for Research, December 22, 1999 and *Voluntary National Test in 8th Grade Mathematics, Test Specifications Outline*, National Assessment Governing Board, March 7, 1998.

[4] 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 TIF was calculated.

- Native English speakers evaluated the similarity of version A to 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.
- MS Word's 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 both 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 (see Appendix A) was used to document any problems that were identified. These judges then met with the two translators to resolve the difficulties.
- MS Word's 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 where 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 dual language 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 sufficient 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; those results are discussed in subsequent sections of this paper.

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 U.S., self-ratings of language proficiency). The dual language 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 Abedi and his colleagues' work (e.g., Abedi, Lord & Plummer, 1997; Abedi, Lord & Hofstetter, 1998). 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 Third International Math and Science Study (TIMSS) and the Program for International Student Assessment (PISA). Three of the administrators were bilingual (two were native Spanish speakers of Mexican descent). The bilingual supervisors led the dual language administrations and the focus groups, while the other field staff led the English-only test sessions. The test administration protocol developed for the dual language study was an adaptation of the NAEP script. We developed a bilingual script for the dual language 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 dual language 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 one point and the constructed response items ranged from one to four points.

Scoring Constructed-Response Items

Our colleagues at NCS took the lead in selecting and training scorers. The 2000 NAEP assessment used bilingual booklets in 4th and 8th grade mathematics tests, and so scorers for the dual language 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.

Inter-Rater Reliability. All constructed-response items were double-scored, in order to obtain a detailed portrait of the types of items (in English or in Spanish) that might cause difficulties in rater agreement. Inter-rater agreement across the two extended and the 13 short constructed-response items was excellent. Using Cohen's kappa and Pearson 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 non-adjacent (i.e., those greater than one 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. Training and scoring took place over a single month, so coder drift was not a factor. As expected, the two extended constructed-response items did produce the largest share of disagreements. NCS coding staff (who are intimately familiar with the NAEP constructed-response items) report that it is especially difficult to train scorers on these two particular items. Because reliability was so high, we chose to use the first score of the pair to compute student test performance.

Study 1: Quantitative Analyses

As described earlier, 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 versus extended time). With this design, we sought to address:

- The speededness of the test, by examining student reached and attempted rates within the standard 45-minute time limit and the 55-minute extended time allocated for this study.
- The degree to which the effects of the dual language 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 psychometric equivalence of the test booklets with one another, by comparing students' mean levels of performance and through differential item functioning (DIF) analyses.

Timing Accommodation

Our study design had built in a within-groups 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 minutes 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-minute marks. Of these 17 students, 12 were in Group A, 4 were in Group B, and 1 was in Group C. Of the 12 Group A students, 5 were sampled to be in the focus groups. These students' use of the additional time might be reflecting 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.

Differential Item Functioning

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. The "thick matching" of ability levels for Mantel-Haenszel DIF analyses, along with the group mean differences reported below, set the conditions for overidentification of DIF. Accordingly, the results of these analyses would not be appropriate for answering questions regarding the psychometric equivalence of the two types of test booklets. We did however, use DIF and classical item analyses in a very circumscribed manner: to select items for the cognitive interview study (Study 3). Although the characteristics of our data made overidentification of DIF likely, we believed that casting a wide net – for the express purpose of selecting items for the cognitive laboratories – was an appropriate decision. Those DIF analyses are described in greater detail under Study 3.

Group Differences

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 non-use of extended time and 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 8th 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 quantitative analyses can be summarized as follows:

- Native English speakers (Group D), on average, outperformed the native Spanish-speakers (Groups A, B and C).
- Group A (native Spanish speakers, less than three years of academic instruction in English, using the dual language booklet) scored significantly lower than the other three groups.
- Group B (native Spanish speakers, three or more years of academic instruction in English, using the dual language booklet) scored significantly lower than Group D.
- Group B and C (native Spanish speakers, three or more years of academic instruction in English, using the English-only booklet) were not significantly different from one another in mean test performance.

Insert Tables 3 and 4 about here

Differences in Adjusted Mean Scores. The next step in our quantitative analyses was a multivariate examination of student performance, where 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 dual language 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, and 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 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.

Insert Table 5 about here

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 dual language test booklet. One way to isolate this effect is to consider the relationship between a student's English language ability and performance, and to see if that relationship 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 dual language booklet condition) and contrasted them to 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 = -.417, n.s.), but found a significant main effect for English proficiency (beta = .175, p = .006), and a marginally significant booklet by English proficiency interaction (beta = .572, p = .06). The interaction pattern was replicated when we examined performance separately for the multiple-choice and the constructed-response sets. The interaction between booklet type and self-reported English proficiency showed students in the dual language booklet condition performed relatively lower than students in the English-only booklet condition, after adjusting for English proficiency.

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 dual language booklet and Group C used the English-only booklet). The patterns that emerged from these analyses were similar to the prior set of regressions, where at every level of English proficiency, students who used the dual language test booklet scored lower than those who used the English-only booklet. These data suggest that a dual language test booklet may not be an aid, and perhaps may even be a hindrance.

Consistent with these findings, Abedi et al. (1997) found a similar type of interaction in their linguistic modification study. They 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 and his colleagues suggested that LEP students perform

best on math tests where the language in which they are tested matches the language in which they are instructed. Their 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 dual language 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. Our own follow-up analyses focused on the language in which students used to answer the test questions. Perhaps the inhibitory effect of the dual language 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 dual language booklet's effect on performance. The preceding analyses were overly simplified in that the comparisons were between students in the dual language 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 dual language 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 dual language test booklet as an accommodation. To investigate this issue, we redistributed the Group A and Group B students into:

- those who chose to answer in Spanish 90% or more of the items (Group AB-Spanish, n=116, see Table 6), and
- those who chose to answer in English 90% or more of the items (Group AB-English, n=77).[5]

Insert Table 6 about here

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'

[5] Five 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.

(i.e., Groups A, B and C) test performance also suggests that the two booklet types are psychometrically equivalent, once English proficiency and language used to answer test questions are accounted for.

Conclusions

Establishing psychometric equivalence between the two types of test booklets posed a challenge because of the sample size. Because we were precluded from a full-fledged DIF analysis, our evaluation of booklet equivalence was limited to a series of regression analyses. The results of the regressions suggest that proficiency in English is an important factor in determining the influence of the dual language test booklet accommodation on students' mathematics test performance. Although there was a weak interaction suggesting that the dual language 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 the quantitative 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 dual language 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 psychometric equivalence between the dual language and English-only test booklets.

The results reported here suggest that implementing additional criteria on top of (and perhaps instead of) the "three years of academic instruction in English" rule merits consideration. States' and other large-scale assessments' inclusion and accommodation policies 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 "three 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.

One final note we would like to make pertains to the motivation of the students in our sample. As our field administrators observed (and as students admitted during the focus groups), our study participants did not apply their best efforts to this test. This low level of motivation must be considered in any interpretations of these data; the relationships observed here may well have been attenuated by the

noise introduced by students' lack of commitment and effort. This caveat holds for the main effects of test booklet type and English proficiency on test performance as well for the interaction between these two variables.

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 dual language test booklet and its perceived utility.

Research Questions

It is important to note that the questions listed below were used as general guidelines for discussion, rather than 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 asking 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 below.

- What were students' perceptions of the quality of the translation used for the dual language test booklet? Did they encounter any unfamiliar phrases or terms? What items were particularly challenging?
- What difficulties did students have with the English-only test booklet? Did they encounter any unfamiliar phrases or terms? What items were particularly challenging?
- How did students use the dual language test booklet? How useful was the dual language booklet?
- 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 six 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 one of the six randomly selected schools, five of the Group A students sampled appeared for the focus group, but only two had signed consent forms. We did conduct the focus group with the two girls, so as not to disappoint them. To compensate for this turn of events, we randomly selected another school and recruited additional Group A students. All told, our 68 focus group participants were distributed in the following manner:

- Group A (less than three years academic instruction in English, administered the dual language booklet): n = 31, sampled from Philadelphia, PA, Los Angeles, CA and Houston, TX.
- Group B (three or more years academic instruction in English, administered the dual language booklet): n = 16, sampled from Deerfield, FL and Philadelphia, PA
- Group C (three or more years academic instruction in English, administered the English-only booklet): n = 21, sampled from Dania, FL and Huntington Park, CA.

Procedures. The focus groups were conducted after the test, while the test-taking experience was still fresh. Students were given a 30-minute 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 in the snacks we had provided.

Focus groups were led by bilingual interviewers and were videotaped. To minimize any self-consciousness that might have arisen from being filmed, we stationed the camera in one spot and simply panned from one student to another, rather than moving about the room. We also spent the first few minutes of the focus group for informal introductions, to acclimate the students to speaking before a camera. At the completion of the focus group, participants were asked to complete and sign a study receipt. Once the receipts were collected, students were given $25 cash each and our thanks for their contributions to the study.

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 the Group C focus group conducted in Huntington Park, CA 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.

Results and Discussion

Prior to reporting the themes that emerged from students' comments, we should note that the focus group experience was a very positive one for these students. The novelty of the event was appealing and they were flattered that the government was concerned enough to ask about their opinions about a test. One student explicitly noted that it was nice that they got chosen to participate in this study, instead of "just the smart kids."

As part of the icebreaker, we elicited a discussion of students' language backgrounds. Group A students (those with fewer than three 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. 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. There were several phrases and words from the test booklet in which we were particularly interested, because these were objects of some debate during the translation process:

- *Papel desdoblado* (unfolded paper). We were concerned that *desdoblado* might not be a familiar term.
- *¿Cuál es el largo de este lápiz aproximando al cuarto de pulgada*? (What is the length of this pencil to the nearest quarter inch?). We thought that *largo* might be ambiguous.
- *Cuestionario* (questionnaire). We were concerned that *cuestionario* might not be a familiar term.
- *La mayoría de 32 niños* (the majority of 32 children). We thought that *mayoría* might be ambiguous.
- *Triángulo ennegrecido* (shaded triangle). We were concerned that *ennegrecido* might not be a familiar term.

We asked students how they interpreted these various phrases, and none had difficulties in defining the words and using the connotations we had intended. One student hesitated when asked about *triángulo ennegrecido* but his pause seemed to be more closely related to an attempt to interpret the diagram (an embedded triangle within a square) than with the phrase itself. The difficulties in understanding expressed by the students appeared to have more to do with familiarity with the mathematical concepts rather than the quality of the translation. Students from each of the three groups (Groups A, B and C) described the words bisector, perpendicular, and parallelogram (whose Spanish versions are virtually identical to the English words) as unfamiliar. One Group A student reported some confusion with the word *promedio* (average). Two Group B students reported suspecting a trick question when they encountered the "none of the above" and "not enough information given" response options.

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 a concern to any of the Group A and Group B students. One Group B student joked about the issue 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 paradoxical 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. Another 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 considered "easy" because did not require them to perform many sub-steps 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, Q29, Carla and Maria's subtraction game, and Session 2, Q29, the broadcast areas of two radio stations) were consistently described as difficult. A Group C student observed that the Carla and Maria subtraction problem was not hard to solve: it was writing the explanation that was the challenge, because in arriving

to the solution, "you just do it." The radio stations item was even more difficult. Students who identified this as a problem item generally said that they simply did not understand what the question was getting at.

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 English-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 fluency, demographic information such as years in the United States). She astutely noted that it would be important to know this sort of information in order 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 comment 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 deleted from the protocol.

Interestingly, although the field supervisors reported that we allocated far too much time for the exam, test length was not much of an issue with the students. Students seemed to possess a slightly fatalistic attitude toward math exams: that all tests were long was a given, and such assessments were events to be endured, as part of the life of a student. Several Group B students said that the thickness of the dual language booklet made the test appear much longer than it actually was. Students from Groups A and B noted with approval our mix of short and long questions.

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 Dual Language Booklet. The dual language 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 comprehension check. One student in Group B noted that this format was particularly useful, 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 understanding. 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 dual language 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 dual language 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 like the FCAT and SAT should offer dual language booklets. Others recommended that we develop dual language tests in all subject areas and to expand our scope to other languages such as Portuguese. We did show Group C students a copy of the dual language test booklet, and that 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 dual language booklet) preferred the dual language 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 dual language 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 dual language 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 dual language test booklet was viewed positively. Although students administered the dual language booklet tended to use one language predominantly, all considered the availability of the second language a benefit. Indeed, we were encouraged to expand dual language testing into the other content areas, as well as to other languages. These findings are an interesting contrast to the quantitative analyses, which indicate that the dual language test booklet has a slight negative effect on the performance of students who have higher levels of English proficiency. But 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 to 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

The objective of the dual language cognitive interview study was to obtain in-depth information on the validity of the translation done on the NAEP 8th grade mathematics items. This subtask addresses the need to gather and implement "systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of [the] language versions" (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, 8th grade students. We closely examined 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 discussed in this section of the report was conducted as a follow-up to Study 1, where we used quantitative analyses to evaluate the psychometric equivalence of dual language and English-only test booklets, and as a complement to Study 2, where we conducted focus groups to obtain information on students' reactions to the test and the test format.

Method

Participants. Eighteen 8th grade students participated in the study. Among the participants, 10 were native Spanish speakers with three or more years of academic instruction primarily in English (usually since first grade). Five were native Spanish speakers with less than three years of academic instruction primarily in English (usually with fewer than three years in the United States), and three were native English speakers. All native Spanish speakers were of Mexican descent. Two of the native English speakers were Caucasian, and the third was East Indian. There were 10 girls and 8 boys in the sample.

To recruit participants, schools and community centers 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.

Selection of Items. Fourteen previously administered NAEP 8th grade mathematics items were selected for the study. These items were among those included in the test administration described previously. Two criteria were used to identify potential cognitive interview items. The first criterion was

related to DIF indicators[6]; we required that an item had to demonstrate one of the following three sub-criteria:

- Significant DIF (at the p = .05 level) for the Group A (native Spanish speakers, less than three years of academic instruction in English, dual language booklet) versus Group D (native English speakers, English-only booklet) comparison
- Significant DIF for the Group B (native Spanish speakers, three or more years of academic instruction in English, dual language booklet) versus Group C (native Spanish speakers, three or more years of academic instruction in English, English-only booklet) comparison

OR

- Very poor discrimination for at least one of the four groups, where "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 sub-criteria. In addition, Item 29 from Session 1 was also included in the initial list of items recommended for the cognitive labs, despite the fact that it did not technically meet any of the three sub-criteria above. 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 non-secure, 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. In order to be included in the set of items recommended for use in the cognitive labs, the item had to have "acceptable" classical item statistics in Group D (native English speakers). We eliminated items that

- were too easy (percent answering item correctly was greater than .95),
- were too hard (percent 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.

Using these sub-criteria, 6 of the 16 items identified through our DIF criterion were judged to have unacceptable classical item statistics. We also excluded an additional item, concluding that it was a

[6] We acknowledge that this combination of "thick matching" of ability levels and large group mean effect size differences set the conditions for over-identification of DIF. However, we cast a wide net for the purpose of selecting items to include in the cognitive laboratory subtask.

borderline item, because it had an incorrect response option that correlated .09 with the total score. This left a total of 10 items for possible inclusion in the cognitive labs. Of these items, six are multiple-choice (MC), three are short constructed-response (SCR), and one is extended constructed-response (ECR).

We decided to include four control items into the cognitive interview protocol. These items had to meet three criteria:

- First, no significant A versus D or B versus C DIF comparisons
- Second, very similar classical item statistics (discrimination indices and difficulties) across groups
- Third, high item-total biserial correlations (greater than .40) in all groups

A number of items met these criteria. However, because of time limitations in the cognitive labs, only four were included in our recommendations for inclusion. 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 dual language 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 dual language 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 semi-structured script for the interview.

Study Design and Sample. This study employed a between-groups design. Eighteen 8th grade students, who differed with respect to their native language and years of instruction primarily in English, were tested under one of two conditions, using either a dual language booklet or an English-only booklet. NAEP inclusion criteria and testing conditions were the independent variables that separated students into four groups (see Table 7). Group comparisons were expected to provide information to help understand DIF results and to illustrate the use, benefits, and drawbacks of the dual language test booklet for students who are native Spanish speakers.

Insert Table 7 about here

Procedures. Prior to interviewing students, interviewers participated in a two-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 two-hour period that included a 15-minute 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 mispronounced 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 I: 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 II: 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 III: Analysis of Student Performance. Interviewers gave detailed reports of a student's performance with the items on data recording forms. They reported the following information:
    - A reading score for each native Spanish speaking student, based on his or her reading of the Spanish language introduction to the items
    - A detailed description of the student's performance on each item, including the think-aloud and the responses to item-specific probes
    - The score (i.e., hit, partial credit, miss) for each item, based on the student's written answer
    - The interviewer's assessment of the student's mastery of the concepts or the skills intended to be measured by each item

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 dual language 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. What factors may explain the differences in psychometric properties for the 10 items identified as showing differential item functioning?
- Research Question 5. Did native Spanish speakers with less than three years of academic instruction in English use the dual language booklet differently from native Spanish speakers with three 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 difficulties interpreting or explaining. In particular, we examined student responses to item-specific probes, which were designed to explore students' understanding of certain words or phrases that seemed potentially difficult to translate.

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).

In order to explore the possible causes of differential item functioning, we took the results of the DIF analyses and examined them in relation to the qualitative data obtained in the cognitive laboratories. The "thick description" of the cognitive processes used by students in solving the problems may uncover potential causes of DIF.

The fifth and final research question addresses the need to provide this type of accommodation (dual language test booklet) for native Spanish speakers with three or more years of academic instruction

in English. The extent to which these students used the Spanish items in the dual language 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 8. The "Responses that Showed Mastery" column was included in order to document our interviewers' judgments whether a student grasped the mathematical content/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 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 three years of academic instruction in English) showed the lowest level.

Insert Table 8 about here

Response Time. We recorded the amount of time that students took to respond to the items in Phase I of the interview, where 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 seconds for each item, SD = 58.0), whereas students in Group A (native Spanish speakers, less than three years instruction primarily in English, dual language booklet) took the most (an average of 137.4 seconds, SD=142.4). Group B (native Spanish speakers, less than three years instruction primarily in English, dual language booklet) and Group C (native Spanish speakers, less than three years instruction primarily in English, English-only booklet) took similar amounts of time to answer each item (Group B mean=103.7 seconds, SD=78.9; Group C mean=109.3, SD=105.8). The large standard deviations in response time for Groups A and C are noteworthy, and are due to students' using 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. 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). Although the phrases were grammatically and syntactically correct, some students were confused about the meaning of the phrases, particularly those with less than three years of academic instruction in English. Changing the verb *voltear* for *girar* could make the item clearer because it better describes the flipping motion of the triangle. The other phrase, "*cuál de las siguientes figuras*," was confusing for one Group A student who was unsure about how many figures she needed to identify (i.e., the possibility of more than one figure). "*Siguientes figuras*" (following figures) seemed to be causing the confusion because it is stated in plural form and it is part of the main question. Rephrasing the question as "*De las siguientes figuras, ¿cuál ...?*" (Of the following figures, which ...?) may convey better the idea that one must choose one out of many.

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 three years of academic instruction in English) seemed to interpret the item differently from 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. As described above, the phrases "*voltear el triángulo de arriba sobre la línea l*" and "*cuál de las siguientes figuras*" created some ambiguity in the item's content to these respondents.

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 three 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 six 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.

Possible Causes of DIF. Results for the 14 items administered during the cognitive interviews are discussed below.

Item M0115401 (how many hours are equal to 150 minutes). DIF analyses showed that native Spanish speakers with less than three years of academic instruction in English using a dual language booklet (Group A) did better on this item than native English speakers with same mathematical ability using an English-only booklet (Group D). However, the cognitive interview data provided no evidence that the dual language booklet format provided an unfair advantage to students in Group A. A discrepancy between item score and mastery existed in one case, where a student in Group A was judged by the interviewer to have mastery of the content, but the student answered incorrectly. Upon probing, this student realized that he failed to correctly represent 30 minutes as a 0.5 hour, as requested by the item. Other students' incorrect responses were due to their lack of mastery of fractions.

Item M016301 (results of flipping a triangle over a line). DIF analyses of this item indicated that booklet type made a difference in the performance of native Spanish speakers with three or more years of academic instruction in English and who were similar in mathematical ability to one another. Group B students (those using the dual language booklet) did better on this item than Group C students (those using the English-only booklet). However, the cognitive interview data provided no evidence that the dual language booklet format provided an unfair advantage to students in Group B.

We did find that some Spanish speakers (one who answered the English version and three who answered the Spanish version) interpreted the item differently from what was intended. Three students in Group A interpreted the phrase "*voltear el triángulo de arriba sobre la línea l*" (flipping the above triangle over the line l) as calling for moving the figure downward, to above line l. Accordingly, these students chose option B instead of the correct response, option E.

Another phrase that was somewhat confusing was "*Cuál de las siguientes figuras*" ("which of the following figures"). One student was uncertain about how many figures she needed to identify. This finding is consistent with what was found in the 1994 Puerto Rico Assessment of Educational Progress, where the use of "*cuál/cuáles*" also tended to be a source of confusion for students (Anderson & Olson, 1996).

Item M022801 (length of one of the longer sides of the rectangle). DIF analyses indicated that native English speakers using an English-only booklet (Group D) did better on this item than native Spanish speakers with less than three years of instruction primarily in English with the same mathematical ability using a dual language booklet (Group A). The cognitive interview data showed that for this item, eight students ignored the phrase "to the nearest centimeter." When they were asked to state the meaning of that phrase, most of them made some kind of gesture (e.g., smiled or acted surprised) and mentioned that they just simply forgot "to round." This may explain the DIF and the low discrimination of the item among students in Group A. Two students in Group A could not explain the meaning of the phrase "*al centímetro más proximo*" (to the nearest centimeter). However, they understood "*redondear.*" Students appear to understand better the verb "*redondear*" than the phrase "*indique a la medida más cercana*" or the verb "*aproxime.*"[7]

Item M022901 (in the number 217, replace the digit 1 with 5). This was one of the four control items that showed no differential item functioning. The cognitive interview data showed that three native Spanish speakers with more than three years of instruction primarily in English used the English version of the item and had a great deal of trouble explaining the intent of the item. Only after reading the item several times were they able to explain what they thought they were supposed to do. The Spanish version of the item also seemed difficult to understand. Three students in Group A, who used the Spanish version, were unable to explain the intent of the item. The fact that several students had problems comprehending the question (two students in Group B mentioned that the item was "weird") suggests that this item should be revised. The use of the passive voice, particularly in a conditional statement (i.e., "be increased if the digit one were replaced"), seemed to be problematic, particularly for native Spanish speakers. Native English speakers showed no difficulty when responding to the item.

Item M023201 (recording a puppy's weight). This was one of the four control items that showed no differential item functioning. The cognitive interview data showed that one native Spanish speaker with less than three years of academic instruction in English answered this item correctly but showed no mastery. He may simply have guessed correctly. When he was asked to explain his answer, he changed it because he assumed that the weight of the puppy would keep increasing 4 pounds every month.

Item M027631 (scale model of a town). DIF analyses of this item indicated that booklet type made a difference in the performance of native Spanish speakers with more than three years of academic instruction in English and who were similar in mathematical ability. Group B students (those using the dual language booklet) did better on this item than Group C students (those using the English-only booklet).

---

[7] The final report of the Puerto Rico Assessment for Educational Progress stated that "redondear" (to round) should not be substituted for "aproximar" (approximate) (Anderson & Olson, 1996).

The cognitive interview data showed that three students (from groups A, B, and C) solved the problem correctly by using general strategies (heuristics) rather than the algorithms required by the item. They did not use division to represent scale models. If they used division at all, they could not explain why. Although the students did not show full mastery, they seemed to have a basic understanding of what could be a plausible answer (e.g., the car is three inches long, thus the house must be at least double that size). They ignored options A and E (which are written in fractions) because they did not seem plausible. They seemed aware that there is a relationship between the numbers mentioned in the item (15/3 = 5; 35/5 = 7), although they were not able to explain how these numbers are related within the context of the item.

Item M027931 (cost of renting a motorbike). This was one of the four control items that showed no differential item functioning. Similarly, no linguistic problems were found in the cognitive laboratories conducted. Native English speakers solved the problem without difficulty, using the algorithm (i.e., algebraic operations) intended by the item. Native Spanish speakers tended to reflect more on what procedures to use. They tended to use a general strategy of substituting numbers within the formula until they obtained a result that matched the value on the table.

Item M028431 (plot the point (5,2) on a grid). DIF analyses of this item indicated that booklet type made a difference in the performance of native Spanish speakers with more than three years of academic instruction in English and who were similar in mathematical ability. Group B students (those using the dual language booklet) did better on this item than Group C students (those using the English-only booklet). DIF analyses also showed that native Spanish speakers with less than three years of instruction primarily in English using a dual language booklet (Group A) did better on this item than native English speakers (Group D) with the same mathematical ability. In the cognitive interviews, the five students who answered incorrectly did so not because of linguistic difficulties, but because they did not remember the notation for identifying points in a coordinate system (i.e., the first value corresponds to the x-axis and the second to the y-axis).

Item M028531 (hair color survey). DIF analyses of this item indicated that booklet type made a difference in the performance of native Spanish speakers with more than three years of academic instruction in English and who were similar in mathematical ability. Group C students (those using the English-only booklet) did better on this item than Group B students (those using the dual language booklet).

However, the cognitive interviews showed no evidence that the use of the dual language booklet made this item more difficult for Group B. Two students (from different groups) received only partial credit because they did not label the portions of the circle with the hair color. One native Spanish speaker with less than three years of academic instruction in English pointed out that he did not know the word

"*cuestionario*" (survey), but he still had enough information to figure out the item's intent. He received only partial credit because he did not label the portions of the circle.

Item M048201 (which figure has the least area). A low biserial correlation indicated that this item discriminates poorly between Group A students (native Spanish speakers with less than three years of instruction primarily in English using a dual language booklet) of differing mathematical ability.

The cognitive interview data suggest that the low discrimination index for this item among students in Group A may be due to the fact that students may define "area" as "perimeter" and still get the right answer. This happened with one student in Group C and two students in Group A. Other students who also confused perimeter with area answered the item incorrectly because they were less careful in their estimates of the length of the sides of each figure. Native English speakers used the proper algorithm to find the correct answer.

Item M049601 (identifying the class president). This was one of the four control items that showed no differential item functioning. The cognitive interview data suggest that there is a possible problem with this item, and students in all groups may have taken advantage of this. Three students (all native Spanish speakers from different groups) concluded that Harriet is the president (the correct response) because she is the only one that is not mentioned in the list of clues. These students were not able to give a more logical explanation (e.g., the other choices did not match characteristics of the president mentioned in the item). It is possible however, that these students had an intuitive understanding that the list was about eliminating possible candidates and that Harriet was the only one who was not eliminated.

Item M051101 (Carla & María's subtraction game). DIF analyses showed that native Spanish speakers with more than three years of academic instruction in English using English-only booklet (Group C) did better than native English speakers with the same mathematical ability using an English-only booklet (Group D).

In the cognitive interviews, students in Groups A, B, and C did not provide answers that gave enough detail to receive full credit. These native Spanish speakers were ambiguous in their written explanations, but gave clear indications that they understood the problem and the right approach to solving it (i.e., the winning player must always have the larger minuend and the smaller subtrahend in the subtraction problem).

Item M055301 (size of the angle formed by the bisectors of two angles). A low biserial correlation indicated that this item did not discriminate between students with high or low mathematical ability in Group A (native Spanish speakers with less than three years of instruction primarily in English using a dual language booklet). The cognitive interviews suggested no clear reason for this low discrimination among Group A students. Knowing the definition of "bisector" determines whether a

student solves this problem correctly. However, one student in this group who lacked full mastery (i.e., the student defined bisectors as "the lines of an angle") selected the correct response. She divided the measure of the right triangle by two because she thought that bisectors were lines in an angle and that lines usually divide things.

Item N263501 (average age of 5 children). DIF analyses showed that native Spanish speakers with less than three years of instruction primarily in English using a dual language booklet (Group A) did better than native English speakers with the same mathematical ability using an English-only booklet (Group D). The cognitive interview data suggest that it is possible that students may select the correct answer through the happenstance application of a series of algorithms. If the first algorithm did not produce a response that matched an answer choice, we witnessed students performing arbitrary mathematical operations until a result that matched a response option was produced. For example, two native Spanish speakers decided to add all the listed numbers. However, the resulting sum (35) did not match any of the response options. So they then decided to divide the result by 5 (i.e., 35/5). Because 7 matched an option, they selected that answer (which was in fact the correct answer). The operations they performed were the correct operations for calculating an average, but they showed no evidence that they knew they were employing the algorithm for calculating an average. Students who marked wrong answers tended to confuse "average" with "median," "mode," and the "midpoint."

Use of the Dual Language Test Booklet by Native Spanish Speakers. Our final research question focused on how Groups A and B used the dual language test booklet. We found that Group B students (native Spanish speakers with more than three 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 three 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 three 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 differential item functioning for items that were administered in a dual language or an English-only test booklet to diverse samples of 8th grade students. A major objective of the study was to identify the extent to which non-equivalent psychometric properties of linguistic versions of the same item can be attributed

to translation issues (e.g., items were not translated accurately or appropriately) or to other confounding factors (e.g., item characteristics).

Despite the limitations of this study (i.e., small n, non-random sample), the rich qualitative data obtained from these cognitive interviews suggest that differences in the psychometric properties of some target items may be 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 different groups who differ with respect to native language and levels of mathematical achievement. The results of the dual language 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 three 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. These students (especially those in Group A) appeared to be testing different strategies that could produce an available (for multiple-choice items) or plausible solution (for constructed-response items). That is, once a multiple-choice option was identified or a reasonable constructed response was developed, students exited this iterative loop. Accordingly, it is possible that the DIF for some of these items is due to the degree to which an item allowed students to identify the correct answer by relying on the application of random heuristics, or on simple "test-wiseness." Of course, in a cognitive interview setting, students could be motivated to try their hardest to answer the items (e.g., social desirability, the anticipation of the honorarium). In the context of a low-stakes test (for which students do not receive scores or other feedback), motivation levels would

undoubtedly be lower. Thus, one cannot state with certainty that lower-performing students will be motivated to use the heuristics they employed in this setting in a real testing situation.

The final issue explored in this study was the use of the dual language test booklet by native Spanish speakers with more than three 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 have taken courses primarily in English since pre-school to those who have received instruction primarily in English only since the 6th grade. The students in this sample used the Spanish items in the dual language test booklet largely as a way to check their understanding of an item, while native Spanish speakers with less than three 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 the quantitative analyses, focus groups and cognitive interviews. Quantitative analyses identified 10 of the 60 items as exhibiting differential item functioning, 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 was consistent with these findings. Students had no difficulties with phrases that had been 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 TIMSS, 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.

There is always room for improvement and we offer two suggestions to those considering a dual language 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. Second, we encourage the continued use of qualitative 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 differential item functioning. But to uncover the reasons behind such difficulties requires input from the examinees themselves, which is best gained through the more qualitative methodologies.

Use and Utility of the Dual Language Test Booklet. Although the test administration data, focus group, and cognitive interview results showed that students tended to use one language predominantly, the dual language 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 dual language 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 dual language 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 dual language format was viewed as more equitable to LEP students who spoke but perhaps were not literate in Spanish, and was considered a good way to help a student learn English.

Review of our regression analyses' interaction plots 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 dual language and English-only test booklets are psychometrically equivalent, 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 psychometric equivalence between the dual language and English-only test booklets, a dual language test booklet appears to be a promising and effective method for inclusion of Spanish-speaking LEP students in large-scale assessments.

Effectiveness and Appropriateness of the Dual Language Format as a Testing Accommodation. Our group comparison analyses were directed toward establishing the effectiveness of the dual language booklet as a test accommodation, and to evaluating the psychometric equivalence of the two types of test booklets. As described previously, initial analyses suggested that after controlling for English proficiency, the dual language test booklet somewhat hindered test performance. However, the follow-up analyses we

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 psychometric equivalency 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 dual language test booklet in Spanish,
- native Spanish speakers who answered the dual language test booklet in English, and
- 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 also must consider whether students actually avail themselves of the accommodation (i.e., answered in Spanish if administered the dual language test booklet). This raises the issue regarding appropriateness of accommodations, and the challenge of accurately identifying students that should receive language accommodations. The basic rule, "three 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 dual language format possibly according an unfair advantage to native Spanish speakers. On the basis of our results, the dual language 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.

Answering the Dual Language Study Research Questions

We now turn to the research questions that guided this study, and close this document by answering each question in turn.

Academic: Are the grammar and language structure used in the native language version correct? According to the judgments of the translation team, the reactions of the focus groups, the data from the cognitive laboratories, the answer to this question is yes. Although a few words and phrases did appear to have some ambiguity within the item context (*voltear*; *el centímetro más próximo*; *cuál de los siguientes*), there did not appear to be any technical or structural difficulties associated with the translation.

Cognitive: Do students understand the native language version of the test questions as a vehicle for assessing mathematics? The answer to this question is derived primarily from the focus group and cognitive interview data. Students in the focus groups regarded the content in the test administered to be appropriate for their grade and with respect to what was being taught in their math classes. To the extent

that students understood math concepts, there did not appear to be any difficulties in the Spanish version of the items. Both the focus group and cognitive interview participants reported difficulties with certain terms (e.g., parallelogram) but those problems seemed to be due to a lack of content knowledge instead of a faulty translation.

Content: Is the content of the native language version of the test questions the same as the English version? As with our response to the academic research question above, the answer to this question is yes, on the basis of the judgments of the translation team, the reactions of the focus groups, and the data from the cognitive laboratories. Although a few words and phrases did appear to have some ambiguity within the item context, the content of the Spanish versions of the items was identical to the English versions of the items.

Cultural: Is the native language version clear and acceptable to the various communities in the United States for whom this is the native language? We specifically asked our focus group participants to react to words and phrases that we anticipated might be ambiguous or unfamiliar, and none had any problems with defining and using the connotations intended. In contrast to other large-scale assessments that adapt their tests into Spanish, we chose the formal (and to some communities, the more proper and respectful) *usted* form of address, rather than the informal *tu*. Students in the focus groups did not seem particularly concerned about this difference. Nonetheless, we recommend prudence and the continued use of *usted*. Therefore, with regard to clarity and acceptability of the Spanish portion of the dual language test, the answer to this question is yes.

Psychometric Equivalence: Is there a psychometric equivalence between the dual language version and the English only version of the test? Results from our test administration data show that yes, the dual language version and the English-only version are psychometrically equivalent – as demonstrated by a lack of significant differences between groups – if the analysis controls for English language proficiency and language used to respond to test items. As discussed above, these findings also suggest that we need to think carefully about how best to identify which students would most benefit from this type of accommodation. The converse should also be considered, as a dual language test booklet might be somewhat detrimental to the test performance of certain types of students (e.g., native Spanish speakers high in English proficiency and whose math instruction is delivered in English).

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 relationships we observed. Nevertheless, the study as a whole (i.e., the results of the regression analyses, focus groups, and cognitive laboratories) should prove valuable to decision-making

regarding inclusion and accommodation policies. Although the results of this study are not definitive, we believe that the data and procedures reported here are also useful in providing guidance to others concerned about providing testing accommodations to English language learners.

References

Abedi, J., Lord, C., & Hofstetter, C. (1998). Impact of selected background variables on students' NAEP math performance (Report No. CSE-478). Los Angeles, CA: Center for the Study of Evaluation.

Abedi, J., Lord, C., & Plummer, J.R. (1997). Final report of language background as a variable in NAEP mathematics performance (Report No. CSE-429). Los Angeles, CA: Center for the Study of Evaluation.

American Institutes for Research (1999). Voluntary National Test in Grade 8 Mathematics Test Specifications. Washington, DC: Author.

Anderson, N.E., Jenkins, F.F. & Miller, K.E. (1996). NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics. Princeton, NJ: Educational Testing Service.

Anderson, N.E., & Olson, J. (1996). 1994 Puerto Rico Assessment of Educational Progress; Technical Report. Princeton, NJ: Educational Testing Service.

Hambleton, R.K. (1996). Guidelines for adapting educational and psychological tests. Paper presented at the joint annual meetings of the American Educational Research Association and the National Council on Measurement in Education, New York, NY. (ERIC Document Reproduction Service No. ED 399 291)

Morgan, D.L. (1988). Focus groups as qualitative research. Newbury Park, CA: Sage.

National Assessment Governing Board (1998). Voluntary National Test in 8th Grade Mathematics, Test Specifications Outline. Washington, DC: Author.

National Center for Education Statistics (1997). 1993-1994 Schools and Staffing Survey: A profile of policies and practices for limited English proficient students: Screening methods, program support and teacher training (Report No. NCES 97-472). Washington, DC: Author.

Organization for Economic Cooperation and Development (1999). Measuring student knowledge and skills. Paris, France: Author.

Olson, J.F., & Goldstein, A.A. (1997). The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress (Report No. NCES 97-482). Washington, DC: National Center for Education Statistics.

Rivera, C., Stansfield, C.W., Scialdone, L., & Sharkey, M. (2000). An analysis of State policies for the inclusion and accommodation of English language learners during 1998-1999. Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

Table 1

Dual Language Test Booklet Study Design

Student Groups Based on NAEP Inclusion Criteria	Dual Language Test Booklet		English-Only Test Booklet	
	Standard Time (45 minutes)	Extended Time (55 minutes)	Standard Time (45 minutes)	Extended Time (55 minutes)
Native Spanish speakers, less than three years of academic instruction in English	Group A			
Native Spanish speakers, three or more years of academic instruction in English	Group B		Group C	
Native English speakers			Group D	

Table 2

Proposed, Sampled, and Obtained Numbers of Students

Student Groups Based on NAEP Inclusion Criteria	**Dual Language Test Booklet** Sampled n: 288 Obtained n: 201 Participation rate: 69.79%	**English-Only Test Booklet** Sampled n: 288 Obtained n: 201 Participation rate: 69.79%
Native Spanish speakers, less than three years of academic instruction in English	Group A Sampled n: 144 Obtained n: 127 Participation rate: 88.19%	
Native Spanish speakers, three or more years of academic instruction in English	Group B Sampled n: 144 Obtained n: 74 Participation rate: 51.39%	Group C Sampled n: 144 Obtained n: 119 Participation rate: 82.64%
Native English speakers		Group D Sampled n: 144 Obtained n: 82 Participation rate: 56.94%

Table 3

Group Mean Levels of Performance

Student Groups Based on NAEP Inclusion Criteria	PERFORMANCE			TEST SCORE RELIABILITY
	MC $\bar{X}$ [SD]	CR $\bar{X}$ [SD]	Total $\bar{X}$ [SD]	
Maximum Possible Score	*45*	*32*	*77*	
Group A (n = 127) Native Spanish speakers, less than three years of academic instruction in English, received DL booklet	17.17 [6.47]	10.09 [5.42]	27.26 [11.05]	.86
Group B (n = 74) Native Spanish speakers, three years or more academic instruction in English, received DL booklet	18.19 [6.95]	12.46 [5.76]	30.65 [11.74]	.87
Group C (n = 119) Native Spanish speakers, three years or more academic instruction in English, received English-only booklet	19.43 [7.70]	13.49 [6.51]	32.92 [13.36]	.90
Group D (n = 82) Native English speakers, received English-only booklet	20.98 [6.88]	14.44 [6.35]	35.41 [12.50]	.90

Table 4

Group Comparisons and Effect Sizes

Comparison	N	Mean	SD	Mean Difference	t-value	Effect Size
Total Test Performance (maximum possible score = 77)						
Group A	127	27.26	11.05	-3.39	-2.05*	-0.30
Group B	74	30.65	11.74			
Group A	127	27.26	11.05	-5.66	-3.60***	-0.46
Group C	119	32.92	13.36			
Group A	127	27.26	11.05	-8.15	-4.94***	-0.70
Group D	82	35.41	12.50			
Group B	74	30.65	11.74	-2.27	-1.20	-0.18
Group C	119	32.92	13.36			
Group B	74	30.65	11.74	-4.77	-2.45*	-0.39
Group D	82	35.41	12.50			
Group C	119	32.92	13.36	-2.50	-1.34	-0.19
Group D	82	35.41	12.50			
Performance, Multiple Choice (maximum possible score = 45)						
Group A	127	17.17	6.47	-1.02	-1.05	-0.15
Group B	74	18.19	6.95			
Group A	127	17.17	6.47	-2.26	-2.50*	-0.32
Group C	119	19.43	7.70			
Group A	127	17.17	6.47	-3.81	-4.05***	-0.57
Group D	82	20.98	6.88			
Group B	74	18.19	6.95	-1.24	-1.13	-0.17
Group C	119	19.43	7.70			
Group B	74	18.19	6.95	-2.79	-2.51*	-0.40
Group D	82	20.98	6.88			
Group C	119	19.43	7.70	-1.55	-1.46	-0.21
Group D	82	20.98	6.88			
Performance, Constructed Response (maximum possible score = 32)						
Group A	127	10.09	5.42	-2.36	-2.92**	-0.43
Group B	74	12.46	5.76			
Group A	127	10.09	5.42	-3.39	-4.43***	-0.57
Group C	119	13.49	6.51			
Group A	127	10.09	5.42	-4.34	-5.29***	-0.75
Group D	82	14.44	6.35			
Group B	74	12.46	5.76	-1.03	-1.11	-0.16
Group C	119	13.49	6.51			
Group B	74	12.46	5.76	-1.98	-2.03*	-0.33
Group D	82	14.44	6.35			
Group C	119	13.49	6.51	-0.95	-1.03	-0.15
Group D	82	14.44	6.35			

Note. Statistical significance is denoted as follows: * $p < .05$; ** $p < .01$; *** $p < .001$

Table 5

Self-Reported Language Proficiency Ratings

Student Groups Based on NAEP Inclusion Criteria	ENGLISH PROFICIENCY		SPANISH PROFICIENCY	
	$\bar{X}$ [SD]	Cronbach's alpha	$\bar{X}$ [SD]	Cronbach's alpha
Maximum Possible Score	*16*		*16*	
Group A (n = 127) Native Spanish speakers, less than three years of academic instruction in English, received DL booklet	10.88 [2.95]	.91	15.22 [1.74]	.86
Group B (n = 74) Native Spanish speakers, three years or more academic instruction in English, received DL booklet	15.15 [1.39]	.83	12.44 [3.17]	.88
Group C (n = 119) Native Spanish speakers, three years or more academic instruction in English, received English-only booklet	14.48 [2.12]	.89	13.02 [3.10]	.90
Group D (n = 82) Native English speakers, received English-only booklet	15.46 [1.76]	.73	8.56 [3.88]	.93

Table 6

Language Used by Native Spanish Speakers to Respond to Test Items

STUDY GROUP	For at least 90% of the items:		Mixed Pattern
	Answered in Spanish	Answered in English	Answered in both Spanish and English
Group A Native Spanish speakers, less than three years academic instruction in English, received DL booklet	109	13	5
Group B Native Spanish speakers, three or more years academic instruction in English, received DL booklet	7	64	3
Group C Native Spanish speakers, three or more years academic instruction in English, received English-only booklet		119	

Table 7

Number of Cognitive Interview Participants, by Demographic Groups and Testing Condition

Student Groups Based on NAEP Inclusion Criteria	Testing Condition	
	Dual Language Booklet	English-Only Booklet
Native Spanish speakers with less than three years of academic instruction in English	Group A (5 participants)	
Native Spanish speakers with three or more years of academic instruction in English	Group B (5 participants)	Group C (5 participants)
Native English speakers		Group D (3 participants)

Table 8

Student Performance and Mastery

Group	Correct Responses	Partially Correct Responses	Incorrect Responses	Total Responses	Responses that Showed Mastery
Group A Native Spanish speakers with less than three years of academic instruction in English using a dual language booklet	27 (40%)	9 (13%)	31 (46%)	67	25 (37%)
Group B Native Spanish speakers with more than three years of academic instruction in English using a dual language booklet	44 (63%)	6 (9%)	20 (29%)	70	46 (66%)
Group C Native Spanish speakers with more than three years of academic instruction in English using English-only booklet	49 (70%)	5 (7%)	16 (23%)	70	53 (76%)
Group D Native English speakers using English-only booklet	37 (88%)	3 (7%)	2 (5%)	42	39 (93%)

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033792

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Study of a Dual-language Test Booklet in 8th Grade Mathematics

Author(s): Duncan, Parent, Chen, Penaranda + Johnson

Corporate Source: American Institutes for Research

Publication Date: April 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Teresa Garcia Duncan	Printed Name/Position/Title: TERESA GARCIA DUNCAN, Sr. Research Analyst
Organization/Address: 1000 Th. Jefferson St. NW, Suite 400 Washington DC 20007	Telephone: 202-295-6853 / FAX: 202-944-5454
	E-Mail Address: tduncan@air.org / Date: 4/4/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)