EDUCATIONAL ASSESSMENT, 9(3&4), 79–105

The Effect of Linguistic Simplification of Science Test Items on Score Comparability

Charlene Rivera
The George Washington University
Center for Equity and Excellence in Education

Charles W. Stansfield
Second Language Testing, Inc.
N. Bethesda, MD

The use of accommodations has been widely proposed as a means of including English language learners (ELLs) or limited English proficient (LEP) students in state and districtwide assessments. However, very little experimental research has been done on specific accommodations to determine whether these pose a threat to score comparability. This study examined the effects of linguistic simplification of 4th- and 6th-grade science test items on a state assessment. At each grade level, 4 experimental 10-item testlets were included on operational forms of a statewide science assessment. Two testlets contained regular field-test items, but in a linguistically simplified condition. The testlets were randomly assigned to LEP and non-LEP students through the spiraling of test booklets. For non-LEP students, in 4 *t*-test analyses of the differences in means for each corresponding testlet, 3 of the mean score comparisons were not significantly different, and the 4th showed the regular version to be slightly easier than the simplified version. Analysis of variance (ANOVA), followed by pairwise comparisons of the testlets, showed no significant differences in the scores of non-LEP students across the 2 item types. Among the 40 items administered in both regular and simplified format, item difficulty did not vary consistently in favor of either format. Qualitative analyses of items that displayed significant differences in *p* values were not informative, because the differences were typically very small. For LEP students, there was 1 significant difference in student means, and

Requests for reprints should be sent to Charlene Rivera, The George Washington University, Center for Equity and Excellence in Education, 1730 N. Lynn Street, Suite 401, Arlington, VA 22209. E-mail: crivera@ceee.gwu.edu

it favored the regular version. However, because the study was conducted in a state with a small number of LEP students, the analyses of LEP student responses lacked statistical power. The results of this study show that linguistic simplification is not helpful to monolingual English-speaking students who receive the accommodation. Therefore, the results provide evidence that linguistic simplification is not a threat to the comparability of scores of LEP and monolingual English-speaking students when offered as an accommodation to LEP students. The study findings may also have implications for the use of linguistic simplification accommodations in science assessments in other states and in content areas other than science.

In recent years, there has been much discussion about how best to assess the school achievement of English language learners (ELLs), referred to in federal legislation as limited English proficient (LEP) students.[1] Two problems faced by those charged with setting inclusion and accommodation policies for state assessment programs designed for system-level monitoring and accountability are (a) the lack of research on the effects of accommodations generally (Shepard, Taylor, & Betebenner, 1998) and (b) the lack of research on how specific accommodations address the linguistic needs of ELLs. This article reports on a study of one accommodation, simplified English, in the Delaware Student Testing Program (DSTP).

Currently, in the Delaware state assessment system, the accommodation of simplifying or paraphrasing test directions or questions is considered a "Condition 3" accommodation. Condition 3 accommodations are not included in school and district means due to concern that students who receive such accommodations will be significantly advantaged over students who do not. However, because the practice of linguistically simplifying test items is considered a promising accommodation strategy for ELLs, an experimental study was designed to assess its effect on the performance of monolingual English speakers who received it. Although the authors realized that the number of LEP students in Delaware was small, they also designed the study to assess the effects of linguistic simplification on the scores of non-LEP students.

The results of the study described in this article should contribute to an understanding of the effects of linguistically simplifying test items on test scores of monolingual English speakers, at least in the context of elementary science assessments in Delaware. The study findings also may have implications for the use of linguistic simplification in science assessments in other states and in subject areas other than science.

---

[1]In this article, LEP students and ELLs are used interchangeably. LEP is the term used in the Elementary and Secondary Education Act to refer to students whose first language is not English and who are designated eligible to receive English-as-a-second-language (ESL) and bilingual services. The term ELL focuses "on what students are accomplishing, rather than on any temporary limitation they face" (LaCelle-Peterson & Rivera, 1994, p. 55).

REVIEW OF LITERATURE ON ACCOMMODATIONS FOR ELLS

Historical Overview

State assessments, like standards-based education, are closely linked to accountability. It is widely believed that school achievement will improve if education systems identify what is to be learned, teach that material, and then assess students on the material to determine the effectiveness of instruction (Committee for Economic Development, 2001). However, concern has been raised about the degree to which standards and accountability systems will include language minority students generally, and LEP students or ELLs specifically (LaCelle-Peterson & Rivera, 1994; Rivera & LaCelle-Peterson, 1993). Rivera, Vincent, Hafner, and LaCelle-Peterson (1997) conducted a survey of state policies during the 1993–1994 school year to discern whether or to what degree states' policies included or exempted ELLs. Responses to a questionnaire sent to state education agencies indicated that 44 of 48 states with state assessment programs permitted ELLs to be excused from one or more state assessments. In 27 of the 44 states, ELLs as a group were routinely exempted from participation in the state assessment program. A key conclusion of the Rivera et al. study (1997) was that if ELLs are to attain the same high performance standards anticipated for native English-speaking students, states must hold ELLs to the same rigorous standards established for their monolingual peers. Rivera and Vincent (1997) recommended the judicious use of accommodations in state assessment programs and in the development of alternate test options; to enable states to document student progress in academic subject areas relative to their English-speaking peers. Rivera and Vincent also recommended that states collect data and conduct studies to evaluate the impact of various types of interventions on LEP student test scores. Subsequently, Rivera and Stansfield (1998) proposed criteria and outlined procedures that can be used to make decisions about the inclusion and accommodation of ELLs in formal assessment programs.

Although Rivera et al. (1997) were conducting their study of state assessment practices relative to ELLs, an independent journalist with support from the MacArthur Foundation conducted an investigation of the testing practices of the 14 largest school districts in the United States (Zlatos, 1994). He found that exemption of the least able students (i.e., those with disabilities, ELLs, and low achievers) was a common practice, and that there was substantial variation in the percentage of students included in large-district testing programs. For example, he found that Philadelphia tested 87% of its students; New York City, 76%; Washington, DC, 70%, and Boston, 66%. The conclusion drawn by Zlatos was that districts use test scores as comparative evidence of the quality of schools without disclosing that the least able students are regularly exempted from participation

in the assessments. Zlatos' findings clearly suggested that learning disabled and LEP students cannot benefit from the standards-based movement unless they are included and their scores reported in state and district assessments and accountability systems.

A subsequent study of state inclusion and accommodation policies for ELLs in the 1998–1999 school year (Rivera, Stansfield, Scialdone, & Sharkey, 2000) showed that states were allowing ELLs to use a variety of accommodations. However, the findings of the study indicated that policies in most states listed accommodations designed for students with disabilities and did not separately list or designate those accommodations responsive to the direct and indirect linguistic needs of ELLs.

Legislative Overview

About the time Zlatos was conducting his investigation, the 1994 reauthorization of the Elementary and Secondary Education Act (ESEA), known as the Improving America's Schools Act of 1994 (IASA), required states to account for ELLs in state accountability systems. It contained requirements that all students reach challenging content and performance standards, and be included in state assessment systems in at least mathematics and reading or language arts.[2] Specifically, the law required that LEP students be assessed annually "to the extent practicable in the language and form most likely to yield accurate and reliable information on what such students know and can do, to determine such students' mastery of skills in subjects other than English" (IASA, Section 1111[b][3][F][iii], U.S. Congress, 1994).

In successive ESEA legislation, the No Child Left Behind Act of 2001 (NCLB; U.S. Congress, 2002),[3] the 1994 requirement to account for ELLs in state accountability systems is clarified and made more stringent. The 2002 law communicates clearly that the inclusion of ELLs and all students in state assessment systems is mandatory. The law specifies that the academic proficiency of all students, including ELLs, must be assessed in reading or language arts and mathematics "not less than once" during grade spans 3–5, 6–9, and 10–12, and, by school year 2007–2008, in science "not less than once" during grade spans 3–5, 6–9, and 10–12 (NCLB, Section 1111, [3][C][v][I-II], U.S. Congress, 2002). By school year 2005–2006, states are to test students yearly in reading or language arts and mathematics in Grades 3–8.

Clearly, to include ELLs in state assessment systems across content areas, a need exists to establish evidence regarding the appropriateness of accommoda-

[2]The rationale for linking standards and assessments is that inclusion of all students in the assessment system will influence what is taught and how it is taught and provide educators with feedback to guide instructional practices.

[3]President George W. Bush signed NCLB into law on January 8, 2002.

tions proposed for these students. The major question to be studied is the effect accommodations render on score comparability, reliability, and validity. This concern, first voiced as part of the IASA legislation, required that all assessment systems used for Title I programs "be valid and reliable and be consistent with relevant, nationally recognized professional standards" (U.S. Department of Education, Office of Elementary and Secondary Education [USDE OESE], 1996). The Draft 1996 IASA Guidance on Assessments stated that

> assessment measures that do not meet these requirements may be included as one of the multiple measures [of adequate yearly progress] if the State includes in its State plan sufficient information regarding the State's efforts to validate the measures and to report the results of those validation studies. (USDE OESE, 1996, p. 15)

Subsequently, guidance issued by the Department reiterated the need for states to utilize valid and reliable assessments (U.S. Department of Education, 1999, 2000). A requirement of IASA was that states create final assessment systems inclusive of all students by the 2001 school year.

Because only a very limited number of in-depth studies evaluating the effects of accommodations on ELL performance, it continues to be critical to study the effectiveness of specific accommodations for ELLs. In addition to the federal legislative impetus begun under IASA and continued and intensified under NCLB, there have been many calls from the education and measurement communities for research to identify appropriate, valid, and reliable accommodations for ELLs. Scholarly publications that address the need for additional focus include the *Position Statement of the American Educational Research Association Concerning High-Stakes Testing in Pre-K–12 Education,* American Educational Research Association, 2000; *Standards for Educational and Psychological Testing,* American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; *Increasing the Participation of Special Needs Students in NAEP: A Report on 1996 NAEP Research Activities,* Mazzeo, Carlson, Voelkl, & Lutkus, 2000; the Teachers of English to Speakers of Other Languages [TESOL] *Position Paper on Assessment and Accountability for ESEA Reauthorization,* TESOL, 2001; and the Office for Civil Rights' guidance on the use of tests to make high-stakes decisions on students, U.S. Department of Education, Office for Civil Rights, 2000. Although research on accommodations for ELLs has begun to be reported at conferences and to appear in the literature (Abedi, Kim-Boscardin, & Larson, 2000; Olson & Goldstein, 1997; Stancavage, Allen, & Godlewski, 1996), studies involving accommodations rarely involve an experimental research design, making it difficult to determine the effects of accommodations on reliability, validity, and score comparability (Shepard et al., 1998).

A Brief History of Simplified English

Simplified English is not a new technique to bridge linguistic gaps across language groups, but rather is historically rooted in the academic and business communities. Ogden (1932) developed the first "Basic English" system to provide a means of cross-cultural communication that would be easy to learn and apply. It consisted of a restricted vocabulary, based on 850 core words, and a restricted grammar system, based on simple sentence structures. Later, Ogden created a dictionary of 20,000 words. In the dictionary, each word was defined using the 850 core words. These included 500 nouns, 150 adjectives, and 100 verbs and other words. However, little attention was given to this innovation in communication (Thomas, Jaffe, Kincaid, & Stees, 1992).

The concept of simplified English was revived in the 1970s and 1980s by companies such as Caterpillar Tractor, and by trade associations such as the aerospace industry associations of Europe and America. In 1972, the Caterpillar Corporation developed a 900-word vocabulary for technical manuals and published *A Dictionary of Caterpillar Fundamental English*. In 1988 the Association Européene de Constructeurs de Matérial Aerospatiale (European Association of Builders of Aerospace Materials) issued a guide for preparing aircraft maintenance manuals called *AECMA Simplified English*. This guide contained a 1,500-word vocabulary and a set of about 40 writing rules focused on style and grammar. Others have further developed and defined "Simplified English," concentrating on refining the core vocabulary (each word with a single unique meaning) and creating glossaries of the technical words specific to the scientific or technical fields in question. Research conducted at Boeing by Shubert, Spyridakis, Holmback, and Coney (1995) on comprehension of two passages from one of Boeing's aircraft maintenance manuals showed that although all readers profited from the simplification, it was nonnative speakers who benefitted the most. This raises the possibility that simplified English might be a promising accommodation for use in testing ELLs.

Research on Simplified English in Testing

We are aware of only four formal studies that specifically examined the effects of linguistic simplification as an accommodation for ELLs. In the first two studies (Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997), the researchers administered simplified mathematics items used on the National Assessment of Educational Progress (NAEP) to eighth graders. In the 1997 study, test booklets containing either a Spanish version, a simplified English version, or the original version of NAEP math items (in regular English) were randomly administered to 1,400 eighth-grade students, half of whom were ELLs in southern California middle schools. Only Hispanic students received the Spanish version. Content experts in linguistics and mathematics rewrote the simplified items at the Center for Research on Evaluation, Standards, and Student Testing. The analyses indicated that

both LEP and non-LEP (fully English proficient, or FEP) students performed best on the simplified version and worst on the Spanish version. Although LEP and non-LEP students performed significantly better on the simplified items, significant differences in item difficulty were obtained on only 34% of the simplified items. Abedi (1997) concluded that linguistic clarification of math items might be beneficial to all students. He noted as well that other factors—such as length of time in U.S., English proficiency, reading competency, and prior math instruction—also had a significant effect on scores.

In their 2001 study, Abedi and Lord simplified the English text for 20 NAEP mathematics items. More than 1,100 students participated, about half of whom were ELLs who took the test containing original and simplified items. No statistically significant differences were found for either group's performance on original or simplified items. Interviews with 36 students revealed that ELLs preferred the modified version of some of the items.

The third study involved the investigation of several accommodations on a math test. Abedi, Hofstetter, Baker, and Lord (2001) administered 35 eighth-grade NAEP math items. Five accommodation conditions were provided: (a) no modification, (b) simplified or modified linguistic structures, (c) a glossary (included definitions of nonmath vocabulary items), (d) extra time, and (e) extra time plus glossary. These conditions were randomly assigned to 946 eighth-grade LEP, non-LEP, FEP, formerly LEP, and a small percentage of monolingual English-speaking students from urban districts in California. A background questionnaire was administered, and all students were given a reading test as a covariate.

Findings of the study indicated that, in general, scores on the reading test correlated with scores on the math test. With regard to the accommodated conditions, all students benefitted from the accommodations. Non-LEP students scored five points higher overall. Although the greatest gains for both groups occurred for glossary plus extra time, extra time, and simplified linguistic structures respectively, the only accommodation that narrowed the score gap between non-LEP and LEP students was linguistic simplification. However, Sireci, Li, and Scarpati (2003) observed that this "narrowing" was due to the fact that the non-LEP students benefitted least under this condition, not that the LEP group did much better than other conditions. Thus, the accommodations studied did not lead to score improvements for the targeted group they intended to help.

In the fourth study (Kiplinger, Haug, & Abedi, 2000), the Colorado Department of Education experimented in 1998 with different versions of released Grade 4 NAEP items from the 1996 NAEP math assessment. They administered a simplified version, a version with an English glossary containing definitions of nontechnical words, and the original version of NAEP items to Special Education, LEP, and regular students at Grade 4. A total of 1,200 students participated in the study. Kiplinger et al. (2000) found no significant difference for the three versions across all students. Neither regular students nor LEP students performed significantly better on either version. However, they attributed this finding to the general diffi-

culty of the test items, which had a mean *p* value of .33. When examining the performance of the students who performed best on the test, they found that this group benefitted most from the glossary and somewhat from the simplified version. They concluded that glossaries and linguistic simplification might benefit all students, and therefore should be used.

Implications

The results of the four studies provide evidence that linguistic simplification of items may have utility as an accommodation for ELLs taking formal assessments. However, more research is needed to attain a full understanding of the effect of linguistic simplification as a test accommodation on the scores of native English-speaking students and ELLs. Only through a full understanding of the effects of linguistic simplification will it be possible to determine whether it should be viewed as a threat to score comparability.

STATEMENT OF THE PROBLEM AND RESEARCH HYPOTHESES

The null forms of the research hypotheses explored in this study are:

1. The mean raw score for Grade 4 and Grade 6 monolingual English-speaking students on linguistically simplified science items will not be significantly greater than that of similar students taking the regular version of the same items on the DSTP Science test.
2. The mean raw score for Grade 4 and Grade 6 LEP students on linguistically simplified science items will not be significantly greater than that of LEP students taking the standard version of the same items on the DSTP Science test.
3. The difficulty of linguistically simplified science items will not be significantly different from the difficulty of regular items for monolingual English-speaking students in Grades 4 and 6 taking the regular version of the same items on the DSTP Science test.
4. The difficulty of linguistically simplified science items will not be significantly different from the difficulty of regular items for LEP students in Grades 4 and 6 taking the regular version of the same items on the DSTP Science test.

INSTRUMENTATION

The DSTP is based on approved content standards for the teaching of English language arts, mathematics, science, and social studies (Delaware Department of Education, Assessment and Accountability Branch, 1999a, 1999b). State assess-

ments in English language arts and mathematics were administered for the first time in the spring of 1998 and again in the spring of 1999 to students in Grades 3, 5, 8, and 10. Assessments in science and social studies for Grades 4 and 6 were field tested in the fall of 1999. The results of the field-testing were used to assemble the final forms of the tests, and the first operational administration occurred in the fall of 2000, when this study was conducted.

In determining which tests to simplify for the study, we examined the sample items available on the Delaware Department of Education (DDOE) Web site. An examination of the sample items in math, science, and social studies indicated that the science items might benefit most from linguistic simplification. This judgment was confirmed through a more detailed inspection of secure math, science, and social studies test items during a visit to the DDOE. Although all the tests contained certain items that could be simplified in terms of the level of language employed, the math test had a lower language load than the science test, and it was determined that the language of the social studies test was more intimately intertwined with the expression of concepts presented and measured on the instrument. Thus, the science test was chosen to be simplified for this study.

Forms of the science assessment at both grade levels consist of 50 items. Thirty-two are four-option multiple-choice (MC) items, and 18 are short answer (SA) items. The MC items are scored dichotomously, as either right or wrong (0 or 1 point for each item). The 18 SA items are scored on a 0–2 scale, with 0 generally representing an inappropriate response or no response, 1 indicating a partially correct response, and 2 representing a fully correct response. To earn a 2, the student must select the correct answer and demonstrate conceptual understanding by explaining why the answer is correct.

RESEARCH DESIGN AND METHODOLOGY

The original plan for the study was to use the full-length operational tests used by the DDOE. However, because the effects of linguistic simplification were unknown, it was feared that those students who took the simplified version would have an unfair advantage. Therefore, it was subsequently decided that the study should be conducted with the field-test items embedded in the operational tests.

Each DSTP science assessment consisted of four forms at each of the four grade levels included within the program. Each form contained a combination of 40 operational items and 10 field-test items. For purposes of this study, two additional forms were created for two grade levels. These additional forms were identical to two of the regular forms, except that the 10 field-test items were simplified. The six operational forms administered at each grade level in the fall of 2000 are listed in Table 1.

All 4th- and 6th-grade students in Delaware participated in the study, because data were collected on all students, regardless of which form students took. Because there were almost 9,000 students at each of these grade levels in Delaware,

TABLE 1
Delaware Student Testing Program
Science Assessment
Forms and Treatments

Science Assessment Form	*Treatment*
A	Regular
B	Regular
C	Regular
D	Regular
E	Simplified
F	Simplified

each form was administered to a sample of almost 1,500 students. The forms were randomly assigned to students through a spiraling procedure, so that in each classroom all six forms were used. Thus, each form was taken by approximately one-sixth of all tested students in the state at that grade level. All forms were randomly assigned to students in every classroom in the state. Random assignment within each classroom assured that the groups that took each form were equal in ability. This full random assignment technique eliminated the possibility that a variation in group scores on the testlets (described in the following) could be due to variation in the ability of students in each group.

Our analyses were based on the test performances of only those students who took forms C through F. Forms A and B did not contain any items that were involved in this study. Forms C and D each contained 10 field-test items as written, reviewed, and revised by Delaware teachers, and then reviewed and edited by test development staff at Harcourt Educational Measurement, the testing contractor used by Delaware. Each item underwent multiple iterations of review and revision both in Delaware and at Harcourt. The 10 field-test items on each form consisted of 6 MC and 4 SA items. Forms E and F contained the 10 field test items included in forms C and D, but in a simplified form.

Twenty items were included in the study at each grade level. The 20 items were divided into two testlets of equal length, with each testlet randomly assigned to monolingual English-speaking and LEP students. In Delaware, MC items are scored as right or wrong, and SA items are scored on a 3-point scale with 0 to 2 points being awarded for each item. All items are based on the state content standards for science. The Grade 4 items assess mastery of the Grades K–3 standards, and the Grade 6 items assess mastery of the Grades 4–5 standards.

The forms were administered to all eligible students. The sample included regular monolingual English-speaking students, some unknown number of FEP bilingual students, and those LEP students who had been in Delaware schools for more than 1 year. Delaware students who have been in the system for less than 1 year are eligible for exemption from participation in the DSTP by state policy. Although a

variety of accommodations are allowed, many LEP students who are tested in Delaware take the tests without accommodations.

Test Simplification

In May 2000, once the DDOE and Harcourt chose field-test items, these were sent to The George Washington University Center for Equity and Excellence in Education team. The six-member team included two middle school science teachers, two applied linguists, and two English-as-a-second-language (ESL) test developers.[4] The intent of the simplification process was to further clarify the task or the context for each item, while reducing its reading difficulty level. The six reviewers met at The George Washington University Center for Equity and Excellence in Education, where digital versions of the tests were projected from a laptop computer onto an LCD screen. This allowed group members to make changes to the file and make iterative revisions that everyone could see. The linguistic simplification process was conducted as follows. As a group, the six reviewers read each item individually. After reading an item, group members reached a consensus about specific science vocabulary, terminology, and structures that needed to be retained in the simplified item to ensure that its original meaning was preserved, and that the construct being measured by the original item was still being tested by the simplified version. They underlined those words and did not subsequently simplify them in any way. The group then examined the item again, this time identifying any difficult syntactic structures or nontechnical vocabulary that was not essential to convey the meaning of the item. Among the language features that were highlighted as potentially problematic for ELLs were passive voice constructions, compound noun phrases, long question phrases, prepositional phrases, conditional clauses, relative clauses, and abstract nouns.[5] In place of those difficult constructions they suggested syntactically simpler alternatives and replaced difficult vocabulary words with higher frequency words that ELLs would have been more likely to have been exposed to in the classroom. Each item was subsequently revised multiple times until the group was satisfied with the result. Once simplified, the field-test items were again reviewed and compared to the original items by DDOE staff to ensure the original meaning of the item had not been altered. The test was then as-

[4]The science teachers were Jim Egenreider and Ray Leonard of Fairfax County, VA, Public Schools. The applied linguists were Dr. Charlene Rivera and Dr. Judith Gonzalez of The George Washington University Center for Equity and Excellence in Education. The ESL test development specialists were Dr. John Miles of the TOEFL test development area at Educational Testing Service and Dr. Charles Stansfield of Second Language Testing, Inc. Miles also coauthored with Rivera and Stansfield (Miles, Rivera, & Stansfield, 2000) a training manual on linguistic simplification of test items that was developed as part of this study.

[5]Some of these linguistic features are discussed in the context of linguistic simplification for ELLs in Abedi, Hofstetter, Baker, and Lord (2001). However, the linguistic simplification procedure described here is not identical to that of Abedi et al. (2001).

sembled and printed. To keep pace with test development timelines established by the test publisher and the state, simplification of the items was completed in a highly concentrated period of time, just 1 day. The simplified test items were sent to Harcourt and the DDOE the following day. The tests were administered between October 10 and October 19, 2000.

The design of the study made it possible to examine a number of issues concerning the science items. These issues relate to the effects of linguistic simplification on regular and LEP students' test scores. The effects could be determined at the level of the testlet (mean score per 10-item testlet by language proficiency status, regular or LEP) and at the level of individual items as well (*p* values by language proficiency status). The study was replicated at two grade levels. Thus, trends and effects at one grade level could be examined for consistency at the other. The design also made it possible to compare the psychometric characteristics (i.e., reliability) of the 10-item testlets by type of item (simplified vs. regular) for each group of examinees at each grade level.

To determine whether there were significant differences in group means across test versions (regular or simplified), *t*-tests for independent samples were conducted at each grade level. Analysis of variance (ANOVA) was used to further explore the data in a way that involved all forms at each grade level simultaneously. The Duncan (1955) and Scheffé (1953) procedures were used to make pairwise comparisons when a significant overall *F* resulted from the ANOVA. Due to the presence of the four SA items (scored 0, 1, 2), ANOVA was also used to compute reliability coefficients for each of the 10-item testlets by form within grade.

RESULTS

A total of 11,306 non-LEP students took one of the eight[6] forms compared in this study. The number of students taking each form was approximately 1,400. Thus, the sample sizes for the non-LEP group were more than adequate for analysis and interpretation. A total of 109 LEP students took one of the eight forms of the test. Because this number was divided across eight forms, the number taking each test form was small, and ranged from 6 to 23 students per form.[7] Because the LEP samples were too small to provide generalizable results, it is not appropriate to try to interpret the findings for the LEP group that participated in this study.

[6]Although six test forms were administered at each grade level, only four contained items that were compared in this study. Therefore, a total of eight forms were compared over the two grade levels.

[7]According to the DDOE, during the 1999–2000 school year only 3% of the students in Delaware were classified as LEP. Among these, many were exempted from participation in the testing program, due to the 12-month exemption policy. In Delaware, all schools use the *Language Assessment Scales* (DeAvila & Duncan, 1975) to identify LEP students.

Item scores, either 0 or 1 for the MC items and 0, 1, or 2 for the SA items, were summed across the 10 regular or simplified items to develop a total score for each examinee. Comparison of means on each type of item (regular or simplified) were made within a grade level for both the non-LEP and LEP groups using both *t*-tests and ANOVA.

t-Test Comparisons

Using *t*-tests within each grade level, the mean of form C was compared to the mean of form E, and the mean of form D was compared to the mean of form F, for both non-LEP (regular) and LEP examinees. The results are displayed in Tables 2 and 3.

Non-LEP examinees. Table 2 shows the mean scores on the regular and simplified items for non-LEP examinees. The difference in fourth-grade students' mean scores on forms D and F was not significant. The difference in mean scores on forms C and E was significant ($p < .05$). The mean score for the sum of the regular items was 6.83, which was significantly greater than the mean of the sum of the simplified items (6.57). However, this difference favoring the regular version is very small, amounting to only 2.5% of the range of possible scores on the testlet.

Table 2 also shows mean score comparisons for the sixth-grade students on forms C and E and forms D and F. For these students, the difference between the mean scores in the two comparisons was not significant at the $p < .05$ level, despite

TABLE 2
t-Test for the Difference Between Mean Scores on the Regular and 10 Simplified Field-Test Items for Grades 4 and 6, Non-Limited English Proficient Examinees

Form	*Type*	*n*	*M*	*SD*	*t*	*df*	*p*	*d**[a]
Grade 4								
C	Regular	1430	6.83	2.64				
					2.63	2840	.009	.025
E	Simplified	1412	6.57	2.69				
D	Regular	1426	6.57	2.65				
					0.42	2840	.676	
F	Simplified	1416	6.61	2.65				
Grade 6								
C	Regular	1415	4.86	2.37				
					0.99	2782	.322	
E	Simplified	1368	4.95	2.24				
D	Regular	1416	6.44	2.60				
					1.66	2837	.096	
F	Simplified	1423	6.61	2.64				

[a]The *d** statistic is discussed in the section Item Difficulty.

the large sample. This finding suggests that there is no advantage for regular English-speaking students who took simplified items when compared to regular English-speaking students who took the regular items. Overall, in three of four comparisons among the non-LEP students, no significant difference in performance was found. In the fourth comparison (Forms C and E at Grade 4), only a very slight difference was found, which favored the regular version.

LEP examinees. Table 3 shows the mean scores on the regular and simplified items for LEP students in fourth and sixth grades. The very small sample size of the LEP groups (*N*s between 6 and 23) strongly suggests that the findings for LEP students cannot be generalized. Among the four comparisons made, only one was statistically significant, and it favored the group receiving the regular items. There were no significant differences at the fourth-grade level. For sixth-grade students, the difference between the mean scores on forms C and E was significant ($p < .05$). The mean for the regular items (4.00) was significantly greater than the mean for the simplified version of these items (2.11). In the other sixth-grade comparison, the difference in means on forms D and F was not significant.

ANOVA

Non-LEP examinees. To further analyze the data for any differences between means, a one-way ANOVA was computed at each of the two grade levels for non-LEP students (see mean scores shown in Table 2). The independent variable, test form, included four forms of the test, C, D, E, and F. The dependent variable

TABLE 3
t-Test for the Difference Between Mean Scores on the Regular and 10 Simplified Field-Test Items for Grades 4 and 6, Limited English Proficient Examinees

Form	*Type*	*n*	*M*	*SD*	*t*	*df*	*p*	*d**[a]
Grade 4								
C	Regular	15	4.67	1.91				
					0.42	31	.677	
E	Simplified	18	4.33	2.52				
D	Regular	23	3.48	1.89				
					1.52	37	.137	
F	Simplified	16	4.38	1.71				
Grade 6								
C	Regular	9	4.00	1.50				
					2.88	16	.011	.025
E	Simplified	9	2.11	1.27				
D	Regular	13	3.23	2.45				
					1.09	17	.289	
F	Simplified	6	2.00	1.79				

[a]The *d** statistic is discussed in the section Item Difficulty.

was an examinee's score on the 10-item testlets. Forms C and D consisted of regular items, and forms E and F contained simplified items. The overall *F* ratio at both fourth- and sixth-grade levels was significant at the $p < .05$ level, suggesting a slight difference in scores across the large sample of non-LEP students.

To determine which means differed significantly from the others, post hoc pairwise comparisons were made using Duncan's (1955) Multiple Range Test. At the fourth-grade level, Duncan's procedure indicated that the mean for form C (regular items) was significantly greater than the means for the other three forms. Scheffé's (1953) more conservative procedure, which keeps the overall error rate at $p < .05$ for all comparisons, failed to show any significant differences between pairs of means.

Post hoc pairwise comparisons at the sixth-grade level provided consistent results with both the Duncan and Scheffé procedures. The means for forms D and F were significantly greater than the means for forms C and E. However, forms D and F are alternate versions of the same 10-item testlet, and on these two versions no differential advantage was found for those examinees who responded to the simplified items when compared to those who responded to the regular items. Therefore, the difference in means was due to a difference in the difficulty of the testlet, rather than in the version of the items that were contained in the testlet. This difference in the difficulty of the testlet was due to the fact that the testlets were constructed from field-test items for which no prior item statistics were available.

LEP examinees. A similar set of ANOVA procedures and post hoc pairwise comparisons was carried out for the LEP examinees. As expected, because of the very small and unequal cell sizes, the overall *F* ratio at each of the two grade levels was not significant.

Reliability Coefficients for the 10-Item Testlets

Reliability coefficients were computed for each of the 10-item testlets by form within grade, as shown in Table 4. ANOVA was used to compute the alpha coefficient due to the presence of four short answer items scored 0, 1, 2. Algebraically, alpha—which is derived by dividing the difference between the mean square between people and the mean square due to residuals by the mean square between people—is identical to KR-20 (Hoyt, 1941).

As shown in Table 4, for the non-LEP group, for a test of this length, the reliability coefficients were quite good. For the fourth-grade sample, the coefficients for the regular and simplified items were (Forms C and E) .50 and .52, and (Forms D and F) .51 and .51. The corresponding coefficients for the sixth-grade sample were (Forms C and E) .60 and .56 and (Forms D and F) .63 and .65. Thus, it appears that for non-LEP students, the reliabilities of the regular and simplified items did not differ.

For the very small LEP group, the reliability coefficients confirmed that the testlets performed inconsistently with this group, with the result that testlets' reli-

TABLE 4
Reliability Coefficients for Non-LEP and LEP Examinees per 10-Item Testlet

N	*Grade*	*Form*	*Alpha*
Non-LEP			
1430	4	C	.50
1426	4	D	.51
1412	4	E	.52
1416	4	F	.51
1415	6	C	.60
1416	6	D	.63
1368	6	E	.56
1423	6	F	.65
LEP			
15	4	C	.19
23	4	D	.23
18	4	E	.55
16	4	F	.00
9	6	C	.00
13	6	D	.75
9	6	E	.02
6	6	F	.63

Note. Non-LEP = non-limited English proficient; LEP = limited English proficient.

ability varied greatly under both conditions (simplified and regular items). For the fourth-grade sample, the range was .00 to .55; for the sixth-grade sample, the range was .00 to .75. For the regular condition, the range was .00 to .75 across the two grade levels. For the simplified condition, the range was .00 to .55 across the two grade levels. This variation is undoubtedly due to the instability of results obtained with small sample sizes. Furthermore, for most of the 10-item testlets, the reliability for the LEP sample was very low. This was due to the short length of the testlets, the very small sample in each LEP group, and the difficulty of the items for the LEP group. Given the probability of getting the items correct by chance guessing, some of the LEP groups who took a particular testlet scored close to chance score. At the sixth-grade level, some of the LEP groups scored at chance level. The low reliability coefficients are a product of the test length, difficulty, the general low ability of the LEP examinees, and the inconsistent measurement inherent in small samples.

Item Difficulty

Because each item was administered in both a regular and a simplified format, it is important to determine whether there is any systematic difference in item difficulty

by item format, and, if so, the magnitude of the difference. Where significant and substantial differences in item difficulty exist, it also is important to examine the items qualitatively, to determine whether there is an apparent reason for this difference. When the cause of such differences can be identified, this information can by used by test developers in future iterations of the test.

The procedure used to determine whether a significant difference exists between the item difficulty (*p* values) for each regular and simplified item requires the construction of 2 × 2 contingency tables to compute a chi-square for each pair of items (see Figure 1). In the case of the SA items, *p* values were determined by summing the percentage of examinees receiving either a 1 or a 2 on the item.

The procedure is equivalent to dividing the difference between two proportions by the standard error of the difference to obtain a normal deviate (*z*), which then can be referred to a table of areas under the normal curve to determine the level of significance. The chi-square procedure is computationally convenient for testing the significance of the difference between two independent proportions. For one degree of freedom, chi-square is equal to the normal deviate squared. The data for each item for non-LEP and LEP student examinees on each form of the test are presented in Tables 5, 6, and 7.

To control for the Type-I error rate in comparisons using multiple *t*-tests to compare testlets and test items, the False Discovery Rate (FDR) procedure was used (Benjamini & Hochberg, 1995). The FDR is the proportion of false negatives among tests for which the null hypothesis is rejected, and the procedure operates on the obtained significance levels to make inferences about a family of comparisons. If the obtained significance level of the difference in the *p* values is equal to or less than *d**, the null hypothesis is rejected; that is, the difference in question remains statistically significant.

Fourth-grade non-LEP examinees. As shown in Table 5, when comparing the *p* values for form C (regular) to form E (simplified), five items (1, 3, 5, 6, 10) had significantly higher *p* values in the regular format, three items (2, 4, 9) had significantly higher *p* values in the simplified format, and two items were not significantly different. For form D (regular) and form F (simplified), no items were significantly different in their *p* values in the two formats. Clearly for the fourth-grade non-LEP examinees, the simplified format was less likely to result in an easier item than the regular format.

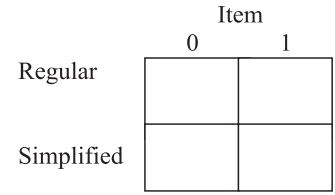

FIGURE 1 Contingency table.

TABLE 5
Comparison of Item Difficulty by Item Condition for Fourth-Grade Non-Limited English Proficient Examinees

Item	*Form C regular p Values*[a] *(n = 1430)*	*Form D regular p Values (n = 1426)*	*Form E simplified p Values*[a] *(n = 1412)*	*Form F simplified p Values (n = 1416)*	*Significance Level*	*d**[b]	*Item Type*[c]
1	*.75*		.59		.01	.025	MC
2	.59		*.66*		.01	.020	MC
3	*.65*		.61		.03	.040	MC
4	.64		*.74*		.01	.015	MC
5	*.60*		.52		.01	.010	MC
6	*.62*		.50		.01	.010	MC
7	.49		.52		.08	.045	SA
8	.53		.56		.09	.050	SA
9	.58		*.62*		.02	.035	SA
10	*.53*		.40		.01	.030	SA
11		.68		.67	.55	.025	MC
12		.55		.57	.26	.010	MC
13		.57		.57	1.00	.050	MC
14		.73		.74	.61	.040	MC
15		.50		.48	.26	.015	MC
16		.57		.58	.88	.045	MC
17		.47		.48	.55	.030	SA
18		.61		.64	.11	.010	SA
19		.63		.64	.56	.035	SA
20		.43		.41	.29	.020	SA

[a]When significant differences occur, the higher of the two *p* values is italicized.
[b]The *d** statistic is discussed in the section Item Difficulty.
[c]MC = multiple choice; SA = short answer.

TABLE 6
Comparison of Item Difficulty by Item Condition for Sixth-Grade Non-Limited English Proficient Examinees

Item	*Form C regular p Values*[a] *(n = 1415)*	*Form D regular p Values (n = 1416)*	*Form E simplified p Values*[a] *(n = 1368)*	*Form F simplified p Values (n = 1423)*	*Significance Level*	*d**[b]	*Item Type*[c]
21	.75		.74		1.00	.050	MC
22	.39		.35		.04	.030	MC
23	.67		.67		1.00	.045	MC
24	.72		*.81*		.01	.020	MC
25	.75		*.80*		.01	.015	MC
26	.36		.40		.04	.025	MC
27	.14		*.19*		.01	.010	SA
28	.18		.17		.45	.045	SA
29	*.39*		.27		.01	.010	SA
30	.28		.30		.28	.035	SA
31		.75		.71	.02	.010	MC
32		.77		.80	.05	.015	MC
33		.80		.80	.93	.050	MC
34		.77		.77	.89	.045	MC
35		.62		*.67*	.01	.010	MC
36		.33		.32	.52	.040	MC
37		.60		.63	.11	.025	SA
38		.73		.75	.16	.035	SA
39		.36		*.41*	.01	.010	SA
40		.21		.24	.11	.020	SA

[a]When significant differences occur, the higher of the two *p* values is italicized.
[b]The *d** statistic is discussed in the section Item Difficulty.
[c]MC = multiple choice; SA = short answer.

TABLE 7
Comparison of Item Difficulty by Item Condition for Fourth-Grade Limited English Proficient Examinees

Item	*Form C regular p Values*[a] *(n = 15)*	*Form D regular p Values (n = 23)*	*Form E simplified p Values*[a] *(n = 18)*	*Form F simplified p Values (n = 16)*	*Significance Level*	*d**[b]	*Item Type*[c]
1	.73		.28		.02	.010	MC
2	.67		.56		.72	.030	MC
3	.53		.50		1.00	.050	MC
4	.40		.50		.73	.035	MC
5	.47		.44		1.00	.045	MC
6	.40		.28		.49	.025	MC
7	.27		.22		1.00	.040	SA
8	.20		.50		.16	.020	SA
9	.13		.50		.03	.010	SA
10	.53		.17		.06	.015	SA
11		.57		.50	1.00	.050	MC
12		.30		.38	1.00	.045	MC
13		.39		.44	1.00	.040	MC
14		.26		.44	.31	.015	MC
15		.39		.19	.29	.020	MC
16		.44		.38	.75	.030	MC
17		.17		.19	1.00	.035	SA
18		.35		.69	.05	.010	SA
19		.22		.63	.02	.010	SA
20		.13		.19	.67	.025	SA

[a]When significant differences occur, the higher of the two *p* values is italicized.
[b]The *d** statistic is discussed in the section Item Difficulty.
[c]MC = multiple choice; SA = short answer.

Sixth-grade non-LEP examinees. As shown in Table 6, for form C (regular) compared to form E (simplified), one item (29) had a significantly higher *p* value in the regular format, three items (24, 25, 27) had significantly higher *p* values in the simplified format, and six items were not significantly different in their *p* values in the two formats. For form D (regular) and form F (simplified), no items had a significantly higher *p* value in the regular format, two items (35, 39) had significantly higher *p* values in the simplified format, and eight items were not significantly different in their *p* values for the two formats. Thus, for the sixth-grade students, neither format was likely to make a difference in item difficulty.

When comparing *p* values for the non-LEP groups, one must keep in mind that a very small difference in absolute value (.03) can produce a statistically significant difference at the $p < .05$ level when analyzing data based on large groups of examinees (see Table 6).

Fourth-grade LEP examinees. As shown in Table 7, for form C (regular) and E (simplified), none of the 10 items were significantly different in their *p* values for the two formats. For forms D (regular) and F (simplified), again none of the 10 items were significantly different in their *p* values for the two formats.

Sixth-grade LEP examinees. For forms C (regular) and E (simplified), one item (11) had a significantly higher *p* value in the regular format, but this difference was not significant when the more conservative *d** statistic was applied. Nine items were not significantly different in their *p* values for the two formats. For forms D (regular) and F (simplified), no items were significantly different in their *p* values for the two formats.

Examination of Simplified Items

In cases where significant differences are found in the difficulty of test items, it is especially important to analyze the changes in wording that were made in the test items. In theory, analysis of the changes to the item will identify the features that make the item easier or more difficult. The features that cause differences in difficulty should show up most clearly in the items where the differences in difficulty are greatest. Thus, we examined the two items that showed the greatest difference in difficulty to see what might have caused the differences. It should be understood that for the non-LEP group, all differences were small. Therefore, even the two most discrepant items show small differences in *p* values.

Grade 4, forms C and E, item 4. The difference in *p* values on the item was .10, in favor of the simplified version (see Table 5). In clarifying the task, a key linguistic principle was used in simplifying the item. Although both the regular and the simplified version of the item included a graphic that illustrated the content

to be tested, only in the simplified version was the examinee told to look at the graphic. In the regular version, it was assumed that the examinee would use the graphic. When the examinee's attention was called to the graphic, the answer became apparent to a greater number of examinees. In the original item, the task to be performed by the examinee was implicit. Linguistic simplification made the related task explicit. This clarification may have helped some non-LEP students who otherwise might not have reacted to the graphic as the test developers assumed they would. In short, the simplification process may have eliminated a weakness in the original item.

Grade 6, forms C and E, item 29. The difference in *p* values on the item was .12, in favor of the regular version (see Table 6). This item also contained a graphic, and the simplified version indicated that the examinee should look at the graphic. However, unlike item 4 in the previous example, the simplified version of item 29 was apparently more difficult. Perhaps the difference is due to other changes in the wording of the simplified version. The simplified version of item 29 avoids the use of the word "consequences" to keep the language simple. However, this word helps convey that the task is to identify the effect of the action introduced in the item. Also, in the simplified version a long stem, in the form of an "if…then" clause, is divided into two sentences. In the simplified item, this also may have reduced the degree to which the item conveys that the examinee is to identify a causal relationship. Thus, at least for the FEP student, "linguistic simplification" may have inadvertently produced a more difficult item.

SUMMARY AND DISCUSSION

When evaluating the efficiency of an accommodation, there are two issues to be determined. First, among those for whom the accommodation is not considered necessary, there is a need to understand whether it provides an unfair advantage to an examinee that receives it over one who does not. Second, if among the first group there is no advantage for those who receive it, then there is a need to understand whether the accommodation actually improves the performance of those who have special needs. And although

> it is difficult to evaluate the effects of accommodations in the context of operational assessment programs, because it is not possible to compare how any given student would have done without the accommodation, … controlled studies are needed to evaluate whether accommodations correct an unfair disadvantage or overcompensate in a way that reduces the validity of assessment results. The ideal study for most accommodations is a 2 × 2 experimental design with both English-language learners

> and native speakers of English being randomly assigned to both accommodated and nonaccommodated conditions. (Shepard et al., 1998, p. 11)

Such a design was employed in this study, and in this case, the accommodation was linguistic simplification. In this study, a team that included experienced test developers, applied linguists, and practicing science teachers linguistically simplified the regular version of items on a fourth- and sixth-grade standards-based state assessment of science. The process was done quickly and efficiently, without delaying the test development timelines. The simplified and regular versions of items were reviewed to ensure meaning had not been altered. The testlets of field-test items were then included on the Delaware operational state assessment. The test forms were assembled so that the field-test portions were identical, except that two of the field-test testlets consisted of regular items and two consisted of the same items in a simplified format. Thus, it was possible to compare the effects of linguistic simplification on item difficulty and student's test performance.

By spiraling the test booklets, the tests were randomly assigned to fourth- and sixth-grade students participating in the DSTP. Separate analyses of the results were completed for regular (non-LEP) students and LEP students for each item condition (regular vs. simplified) and for each grade level. The results were broken down by total score on the testlet and by item difficulty (p value).

Only a small number of LEP students participated in the 2000 Delaware state assessment. Therefore, as expected, few significant differences were found in the LEP analyses, and it is not possible to draw any conclusions from the results regarding the effects of the simplified items on LEP students. However, the samples for the regular FEP students were quite large, with the result that conclusions can be drawn based on the data.

The results of the study support the conclusion that among FEP students, linguistically simplified items are normally of no help to students taking a test. That is, as a test accommodation, linguistically simplified items function like eyeglasses. If a student does not need eyeglasses to see clearly, then the glasses do not improve his or her vision. On the other hand, if a student has deficient vision, then glasses will improve vision. Thus, when taking a test, glasses level the playing field for those who need them, so that everyone is able to see with an adequate degree of clarity, while not giving those who use glasses an advantage over those who do not.

In this study, there was no significant difference in the mean raw scores of English-speaking students who took simplified testlets and those who took the same testlets with regular wording.[8] This is an important finding, because it shows that lin-

[8]That is, for three of the four comparisons among the non-LEP students, no significant difference in performance was found. In the fourth comparison (forms C and E at Grade 4), only a very slight difference in favor of the regular version (amounting to only 2.5% of the range of possible scores on the testlet) was found.

guistic simplification can be used without fear of providing an unfair advantage to those who receive it, and thereby affecting the comparability of scores across examinees obtained under this condition. With this knowledge in hand, educational testing specialists concerned with the identification of ways to meaningfully include more students in an assessment program can offer linguistically simplified science assessments to LEP students without fear of providing them with an unfair advantage that would invalidate their scores. Because linguistic simplification is able to reduce the language load, it is likely that it can reduce the role of language proficiency in achievement test scores, generally. Because language is not the construct tested in science tests, then reducing the role of language should reduce the amount of construct irrelevant variance in test scores, particularly for LEP students.

Other studies should now address this issue of the usefulness of linguistic simplification for LEP students taking formal and high-stakes assessments. If experimental studies involving large samples of LEP students who are randomly assigned to treatments show that the LEP students who receive simplified items perform statistically and meaningfully better than those who receive the regular, unsimplified version of such items, then the utility of linguistic simplification in meeting the needs of LEP test takers will be established. "If assessment accommodations [work] as intended, the results should show an interaction effect. The accommodation should improve the performance of English-language learners but should leave the performance of native-English speakers unchanged" (Shepard et al., 1998, p. 11). Such studies will have to take place in states or large districts that have large numbers of LEP students. Then, before the decision to alter assessment conditions is made, validity data must be collected and carefully analyzed.

In this study, we chose to simplify science test items after examining Delaware tests in math, science, and social studies. We noted that the science assessments involved a greater language load than the math assessments, but less than the social studies tests. We also assumed that it would be harder to linguistically simplify the social studies assessments without affecting the clarity with which concepts were communicated and without interfering with the use of terminology that is central to the field of social studies. Because the items in the math assessment had a lower language load than those in the science assessment, we supposed that any effect that might result from linguistic simplification would be more evident on the science tests. Nonetheless, we were unable to find any systematic effect on the science tests. Future research should examine the effects of linguistic simplification on formal assessments in other subject areas, such as math and social studies.

Although it is unfortunate that the study could not address the effectiveness of linguistic simplification for ELLs, the study was successful in showing that tests and items can be linguistically simplified without compromising score comparability, at least in the area of science. Of course, the process of linguistically simplifying test items requires appropriate expertise, and it must be carried out with care. The result of the process of linguistic simplification must be to

make the items accessible to ELLs without altering the difficulty of the content. At times, language and content interact; in these cases, it is not possible to linguistically simplify items without simplifying the content. Because the simplification process must be managed with caution, like item writing in general, it cannot be assumed that all linguistic simplification efforts will achieve the same result. However, if a future study demonstrates that linguistic simplification is effective for ELLs, additional research efforts will need to identify the linguistic features of items that cause problems for ELLs, and the procedures to be observed and the linguistic or organizational features to be implemented in the revision of test items. We encourage others to pursue this promising avenue for future research involving the testing of ELLs.

ACKNOWLEDGMENTS

In June 1999, the state of Delaware issued a request for proposals for research on accommodations for LEP students. The authors responded to that request with a proposal to carry out this study of the accommodation of linguistically simplifying test items by carrying out an experimental study of the effect of simplifying science test items. We express our appreciation to the DDOE for having funded this study. At the DDOE, Drs. Wendy Roberts, Nancy Maihoff, and Liru Zhang were helpful and supportive, as was Harcourt Educational Measurement, Delaware's contractor for the science test. We also acknowledge the contributions of Dr. John Martois, an independent statistical consultant, who prepared the data and carried out the statistical analyses reported in this study. Finally, we appreciate the contributions of Melissa Bowles of Second Language Testing, Inc.

An earlier version of this article was presented at the annual meeting of the American Educational Research Association, Seattle, Washington, April 12, 2001.

REFERENCES

Abedi, J. (1997). Impact of selected background variables on students' NAEP math performance. Los Angeles: Center for the Study of Evaluation. Draft deliverable.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Kim-Boscardin, C., & Larson, H. (2000). *Summaries of research on the inclusion of students with disabilities and limited English proficient students*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219–234.

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, Center for the Study of Evaluation.

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in pre-K–12 education. *Educational Researcher, 29*(8), 24–25.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Association Européene de Constructeurs de Matérial Aerospatiale.(1988, January). AECMA Simplified English Document: A guide for the preparation of Aircraft maintenance procedures in the international aerospace maintenance language. (Issue 1, Revision 4, AECMA Document No. PSC–85–16598).

Benjamini, Y., & Hochberg, Y. (1995). Controlling for the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, 57,* 289–300.

Committee for Economic Development. (2001). *Measuring what matters: Using assessment and accountability to improve student learning.* New York: Author.

DeAvila, E., & Duncan, S. (1975). *Language assessment scales.* Monterey, CA: CTB-McGraw Hill.

Delaware Department of Education, Assessment and Accountability Branch. (1999a, January). *Delaware Student Testing Program: State summary report, 1998 administration.* Dover, DE: Author. Retrieved October 21, 2003 from http://www.doe.state.de.us

Delaware Department of Education, Assessment and Accountability Branch. (1999b, March). *Guidelines for the inclusion of students with disabilities and students with limited English proficiency.* Dover, DE: Author.

Duncan, D. B. (1955). Multiple range and multiple *F* tests. *Biometrics, 11,* 1–42.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 53–160.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, June). *A math assessment should assess math, not reading: One state's approach to the problem.* Paper presented at the 30th National Conference on Large Scale Assessment, Snowbird, UT.

LaCelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64,* 55–75.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities.* Washington, DC: National Center for Education Statistics.

Miles, J. E., Rivera, C., & Stansfield, C. W. (2000). *Leveling the assessment'playing field': Making science test items accessible to English language learners.* Arlington, VA: George Washington University, Center for Equity and Excellence in Education.

Ogden, C. K. (1932). *Basic English, a general introduction with rules and grammar.* London: Paul Treber & Co.

Olson, J., & Goldstein, A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress.* Washington, DC: National Center for Education Statistics.

Rivera, C., & LaCelle-Peterson, M. (1993). *Will the national education goals improve the progress of English language learners? ERIC Digest.* Washington DC: ERIC Clearinghouse on Language and Linguistics. (ERIC Document Reproduction Service No. ED362073)

Rivera, C., & Stansfield, C. W. (1998). Leveling the playing field for English language learners: Increasing participation in state and local assessments through accommodations. In R. Brandt (Ed.), *Assessing student learning: New rules, new realities* (pp. 65–92). Arlington, VA: Educational Research Service.

Rivera, C., Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during*

1998–1999. Arlington, VA: George Washington University, Center for Equity and Excellence in Education.

Rivera, C., & Vincent, C. (1997). High school graduation testing: Policies and practices in the assessment of English language learners. *Educational Assessment, 4,* 335–355.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). *Statewide assessment programs: Policies and practices for the inclusion of limited English proficient students*. *ERIC Digest*. Washington DC: ERIC Clearinghouse on Measurement. (ERIC Document Reproduction Service No. EDO–TM–97–02)

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40,* 87–104.

Shepard, L. A., Taylor, G. A., & Betebenner, D. (1998). *Inclusion of limited English proficient students in Rhode Island's grade 4 mathematics performance assessment* (CSE Tech. Rep. No. 486). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Shubert, J. K., Spyridakis, J. H., Holmback, H. K., & Coney, M. B. (1995). The comprehensibility of simplified English in procedures. *Technical Writing and Communication, 25,* 347–369.

Sireci, S. G., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature.* (Center for Educational Assessment Research Report No. 485). Amherst: University of Massachusetts, School of Education.

Stancavage, F., Allen, J., & Godlewski, C. (1996). Study of the exclusion and assessability of students with limited English proficiency in the 1994 Trial State Assessment of the National Assessment of Educational Progress. In National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment, *Quality and utility: The 1994 trial state assessment in reading* (pp. 172–175). Stanford, CA: National Academy of Education.

Teachers of English to Speakers of Other Languages Elementary and Secondary Education Act Reauthorization Task Force. (2001). Board endorses position papers for ESEA reauthorization effort. *TESOL Matters, 11*(1), 1, 4.

Thomas, M., Jaffe, G., Kincaid, J. P., & Stees, Y. (1992). Learning to use simplified English: A preliminary study. *Technical Communication, 39,* 67–73.

U.S. Congress. (1994). *Improving America's Schools Act of 1994.* Public Law 103–383. Washington, DC: Government Printing Office.

U.S. Congress. (2002). *No Child Left Behind Act.* Public Law 107–110. Washington, DC: Government Printing Office.

U.S. Department of Education. (1999, November). *Peer reviewer guidance for evaluating evidence of final assessments under Title I of the Elementary and Secondary Education Act*. Washington, DC: Author.

U.S. Department of Education. (2000, July). *Summary guidance on the inclusion requirement for Title I final assessments*. Washington, DC: Author.

U.S. Department of Education, Office for Civil Rights. (2000). *The use of tests when making high-stakes decisions for students: A resource guide for educators and policymakers*. Washington, DC: Author.

U.S. Department of Education, Office of Elementary and Secondary Education. (1996, October). *Title I, Part A Policy Guidance: Improving basic programs operated by local educational agencies. Guidance on standards, assessments, and accountability.* Washington, DC: Author.

Zlatos, B. (1994). Don't test, don't tell. *American School Board Journal, 181*(11), 24–28.