

Do Proper Accommodation Assignments Make a Difference? Examining the Impact of Improved Decision Making on Scores for English Language Learners

Rebecca J. Kopriva, *University of Wisconsin*, Jessica E. Emick, *Inner Health Ministries*, Carlos Porfirio Hipolito-Delgado, *University of Colorado*, and Catherine A. Cameron, *Educational Researcher*

Does it matter if students are appropriately assigned to test accommodations? Using a randomized method, this study found that individual students assigned accommodations keyed to their particular needs were significantly more efficacious for English language learners (ELLs) and that little difference was reported between students receiving incomplete or not recommended accommodations and no accommodations whatsoever. A sample of third and fourth grade ELLs in South Carolina (N = 272) were randomly assigned to various types of test accommodations on a mathematics assessment. Results indicated that those students who received the appropriate test accommodations, as recommended by a version of a computerized accommodation taxonomy for ELLs (the selection taxonomy for English language learners accommodations; STELLA), had significantly higher test scores than ELLs who received no accommodations or those who received incomplete or not recommended accommodation packages. Additionally, students who were given no test accommodations scored no differently than those students that received accommodation packages that were incomplete or not recommended, given the students' particular needs and challenges. These findings are important in light of research and anecdotal reports that suggest a general lack of systematicity in the current system of assigning accommodations and a tendency to give all available accommodations regardless of individual child characteristics. The results also have important implications for how future accommodation research should be structured to determine the benefits of particular accommodations and accommodation packages. This study would suggest that control and treatment groups should be assembled based on specific student needs in order for direct comparisons to be made.

Keywords: accommodation decisions, accommodations, English language learners, STELLA

Validity of Accommodation Assignments

With the increased emphasis placed on standardized testing for measuring student achievement (e.g., Improving America's Schools Act of 1994; No Child Left Behind Act of 2001), there is simultaneously a call for the inclusion of more diverse student groups, including English language learners (ELLs), in standardized testing. Questions of how to incorporate ELLs into large-scale academic accountability systems have yet to be completely answered, although the usage of testing accommodations has been viewed as a key method to meaningfully incorporate ELLs into these assessments (Butler & Stevens, 2001; Rivera & Stansfield, 2003). Unfortunately, however, research on testing accommodations has yielded mixed results regarding their overall effectiveness with ELLs (Abedi, 2001; Abedi, Courtney, & Leon, 2003; Emick, Kopriva, Chen, Mislevy,

Rebecca J. Kopriva is a Visiting Research Scholar at University of Wisconsin Madison, 1025 W. Johnson St., Madison, WI 53706; rkopriva@wisc.edu. Jessica E. Emick is a Postdoctoral Fellow at Inner Health Ministries, 17730 Lorain Road, Cleveland, Ohio 44111. Carlos Porfirio Hipolito-Delgado is an Assistant Professor at the University of Colorado at Denver and Health Science Center, Campus Box 106, Denver CO 80217. Catherine A. Cameron is an independent researcher in Bethesda, MD.

& Carr, 2006; Mann, Emick, Cho, & Kopriva, 2006).

Much of this research has been conducted using test accommodations with ELLs that has used a blanket method of accommodations—the usage of the same set of one or more accommodations with all ELLs in a group regardless of individual characteristics (i.e., different levels of English Language Proficiency (ELP) or proficiency in their primary language [L1]). Almost no research has been done to investigate the effectiveness of specific accommodations for those students who need them (versus those who do not). Further, little research has been done to examine the validity of accommodation assignment—specifically does validity improve if accommodations are assigned systematically based on individual characteristics? The method of providing blanket accommodations to ELLs in research and practice has come under criticism (e.g., Hofstetter, 2003). Butler and Stevens (2001) have also called for a more systematic approach to provide test accommodations to ELLs. This issue was clearly brought to light with the 2002 complaint filed with Office of Civil Rights against the Pennsylvania Department of Education (PDE) charging that the PDE's method of assigning accommodations was inadequate to demonstrate assessment validity for ELLs.

We do not dispute PDE's view that the provision of accommodations, carefully tailored to the needs of the student and the demands of the test, can sometimes improve the validity of tests given in English to English language learners. But fitting the accommodations to the student and the test is a crucial part of the process. Some accommodations, in particular situations, will raise an ELL student's scores more effectively than other accommodations. Some accommodations will be appropriate for one student but not for another. Some accommodations, under certain circumstances, will actually give a student an unfair advantage. PDE's haphazard approach, which includes no guidance to schools and no way of ensuring that teachers make informed decisions, is inadequate to ensure that accommodations are correctly tailored to students' English proficiency and needs. (Education Law Center, 2002, p. 17)

Heeding this call, school districts and states have continued to refine their accommodation policies. Unfor-

tunately, however, many of the educational agency practices are not based in research, and, surveys have shown, admissible and inadmissible accommodations have been inconsistently identified across the states (Samuelson & Kopriva, 2004). With the exception of very broad indicators, for the most part a systematic method of assigning test accommodations for individual students ELL students has not been conceptualized to date, much less utilized to make recommended accommodation decisions for individual students.

After a discussion about the implications of properly assigning accommodations, the results of a study that addresses these issues will be presented. The goal of the study was to determine if the validity of the students' performance on a mathematics test varied as a result of appropriately or inappropriately assigning accommodations to participating Spanish-speaking students. The randomized assignment of treatment conditions was designed so that, without intervention, performance should be similar across groups, and this approach would therefore allow the researchers to make direct comparisons about the influence of accommodation assignments that were identified as (1) recommended, (2) incomplete or not recommended, or (3) absent. On the basis of prior research and theory, it was hypothesized that appropriate assignment of accommodations would result in higher performance relative to the other two groups. It was argued that, because of the randomized design, a boost would indicate increased validity for the group of students impacted by the boost. These results will be examined in light of the purpose and individualized approach encapsulated in *STELLA* (the selection taxonomy for English language learner accommodations), a multisource, computerized system that assigns accommodations based on the needs of each student (Kopriva, 2005; Kopriva & Koran, 2006).

Need for Systematic Assignment of Test Accommodations with ELLs

The limited English proficiency of many ELL students raises questions about whether standardized tests yield a valid assessment of the knowledge of ELLs (e.g., Butler & Stevens, 2001; Castellon-Wellington, 2000; Hofstetter, 2003; Kopriva, 2002). It might be argued that when typical standardized academic assessments are administered to ELLs—especially those with lower En-

glish language proficiency—the assessment to some extent becomes an evaluation of language and cultural skills rather than the actual content area the assessment was designed to measure (Kopriva, 2000). This point is reinforced by a number of studies with ELLs where students' English language proficiency was significantly related to their performance on a standardized examination (Castellon-Wellington, 2000; Hafner, 2001; Hofstetter, 2003), and a lack of systematic validity studies to determine when accommodation scores are producing valid inferences for individual ELL students (Koenig & Bachman, 2004).

Prior to the mid-1990s, the concern of validity of the assessment inferences led to ELLs being routinely excluded from standardized achievement of assessments (Butler & Stevens, 2001). More recent legislation called for the incorporation of ELLs into statewide standardized assessment (Improving America's School Act, 1994; No Child Left Behind Act, 2001) because it was viewed that the exclusion of ELLs from standardized assessment clouded whether or not schools, local education agencies (LEA), and state education agencies (SEA) were appropriately evaluating the learning of all students. To the extent that this population was properly included in statewide assessments, it was argued that the growth of ELLs' academic achievement could be systematically evaluated across schools, separately and in comparison to other groups. In order to support the learning of ELLs and others, targeted policies based on disaggregated but comparable results could be implemented and evaluated (Abedi et al., 2003; Lara & August, 1996).

Multiple methods that attempt to properly include students in large-scale academic tests have been utilized. In some cases, translated versions of standardized tests have been created, alternative testing conditions have been employed, and, most often, test accommodations have been identified and adopted (Butler & Stevens, 1997; Rivera & Stansfield, 2003). However, given the plethora of languages spoken by ELLs, the limited literacy of many students in their L1, the difficulty of producing appropriately translated tests, English-only legislation, and movement to test students in the language of instruction, it appears that it is often impractical to produce and use a wide range of written test translations.

A number of approaches, such as parallel plain language forms, use of L1 aids, and pretest supports are currently being examined (Emick, Monroe, Kopriva, & Sprehn, 2007; Kopriva, Winter, Wiley, Emick, & Chen, in press). Recently, some states and other agencies have responded by attempting to address the needs of early ELLs by exploring non-standard methods of assessment but, to date, the technical rigor of many of these for large-scale use has yet to be adequately demonstrated (S. Rigney, personal communication, May 8, 2006).

In part, because of the difficulties inherent in widely developing and implementing alternative approaches, many researchers and practitioners have suggested that appropriate test accommodations could be identified and utilized in the statewide tests to validly include many ELLs (Butler & Stevens, 2001). Depending on state policy and practical applications, these accommodations should diverge from typical accommodations used for students with disabilities and take into consideration L1 proficiencies as well as other issues and strengths unique to this population. However, the current status of the accommodation research has, to a large extent, hampered the ability of states to meaningfully identify and assign accommodations. For instance, Hafner (2001) and Rivera and Stansfield (2003) found that the test accommodations significantly improved test performance of some ELLs. On the other hand, Castellon-Wellington (2000), Liu, Anderson, Swierzbis, Spicuzza, and Thurlow (1999), and Albus, Thurlow, Liu and Bielinski (2005) did not find a difference in test performance between students who received test accommodations and those students that did not receive accommodations.

If accommodations are to be utilized as a primary approach to include a large percentage of English language learners, it appears that proper assignment of accommodations based on student need may address some of the shortcomings and confusion highlighted in the research. Also, it seems that systematic needs-based assignment of accommodations should help produce clearer guidance to the field about which accommodations are useful and for whom. For instance, Hofstetter (2003) found that testing students in the same language in which they were instructed significantly improved the performance of ELLs. She found that those ELLs

who received instruction in English performed better on a plain language English exam and those students who were instructed primarily in L1 performed better on an L1 exam. This finding not only points to the importance of differentiated assignment of test accommodations based on individual characteristics, but also suggests that the ELP and L1 levels of students are important to consider when providing test accommodations to English learners.

While many states do have policies regarding allowable accommodations, most SEAs do not have specific criteria or routine trainings that explain which accommodations should be given to which students (Albus, Thurlow, Liu, & Bielinski, 2005; Kopriva, Koran, & Hedgspeth, 2007). Analogous to what was identified in Pennsylvania (Education Law Center, 2002), state and local districts, in general, do not seem to provide definitive guidelines, empirical or otherwise, about matching certain accommodations to certain child characteristics. Further, in a review of policies across the states, *STELLA* researchers found little coordination between the individuals who would actually be making the accommodation assignments and the policy makers, or guidance about how policies were to be implemented (Kopriva & Hedgspeth, 2005). Rivera and Collum (2006) also noted that decision makers lack definitive guidelines for assigning accommodations, resulting in a sometimes arbitrary connection between criteria and accommodation decisions.

In many cases, the ultimate decision about which accommodation(s) an ELL student receives on a large-scale academic assessment is left in the hands of the student's teacher or local ELL specialist (Liu et al., 1999). These researchers found that assignment of accommodations to ELLs has largely been based on anecdotal information and in some cases it is not clear how these decisions are made. Adelman (1992) reported that experts are generally unaware and unable to describe their exact decision-making process but can verbalize which process is best. However, teacher focus groups conducted at the outset of the development of *STELLA* found that it was difficult for teachers to explain their assignments (Douglas, 2005). Recent evidence suggests that, even with some training, leaving accommodation deci-

sions to the discretion of these experts alone may be problematic. Two reports found that teachers or SEA specialists struggled to differentiate accommodations for individual students even though they recognized that these students differed widely on salient variables (Plake & Impara, 2006; Koran, Kopriva, Emick, Monroe, & Garavaglia, 2006). For instance, Koran and others found that teachers tended to recommend a similar set of accommodations for most of their students even though they were able to purposely identify students with diverse needs. It is surmised that part of this confusion may stem from a teacher's dual roles, as student advocate and as implementer of policy (Douglas, 2005).

While teachers or specialists may not be able to clearly differentiate recommended accommodations, their "fall back" approach often seems to be to assign all possible accommodations. The early *STELLA* teacher focus groups examined what accommodations were typically used in classrooms, what types of information were used by the teachers to make the decision about how to use them with individual students, and how the teachers made decisions about what accommodations to assign to their students for large-scale achievement testing purposes. A prevalent attitude that became apparent, especially with regard to the large-scale testing, was that more accommodations were better. See the following example.

Moderator: What if it was a state assessment, and pretend the sky is the limit in what you can do to accommodate the student.

Teacher 4: The sky is the limit on what I could accommodate? Then I would give them everything that I could give them, I would give them the bilingual dictionary that they need, I would give them whatever.

Out of concern for the best interest of their students, the teachers, again and again, suggested that they would encourage the large-scale testing coordinators to provide all allowable accommodations for their students, regardless of need. It is not surprising, then, that a reasonably large number of ELL students may be receiving inappropriate sets of accommodations when they take their statewide academic assessments.

Clearly, the decisions regarding which ELLs should receive a particular accommodation are not easily

reached. Whether the impasse is because of the lack of clear sets of guidelines or proper training, the difficulty of educators to explain or implement methodical decision-making processes across students with like needs, or because of the dual roles of educators, the result is that students may not be getting what they need when accommodations are assigned. Based on these concerns there appears to be a need for a more systematic method of assignment, and one that, most likely, includes salient information from more than one source.

Addressing the Issue

Recently Rivera and Collum (2006) introduced an *ELL-Responsive Taxonomy*, which emphasizes the individual needs of ELLs and provides accommodation recommendations based on two broad categories (direct or indirect linguistic support). The current study used some of the decision-making algorithms encapsulated by the *STELLA* system. As noted above, the *STELLA* system provides a systemic, individualized, and computerized method for deciding the particular types of accommodations a student should receive on achievement tests, and is based on obtaining information along seven dimensions. *STELLA* was developed based on a series of large-scale and classroom accommodation reviews, decision-making reviews, and formative studies (i.e., teacher focus groups, parent and teacher interviews, and expert panels), and subsequently evaluated by this and another study (Koran & Kopriva, 2006). *STELLA* collects data on individual students from three sources: cumulative school records, teachers, and parents. These three sources provide independent assessments of a student's familiarity with standardized assessments as administered in U.S. schools, English language proficiency levels (ELP), L1 proficiency levels, cultural proximity of past schooling experiences to current U.S. schooling experiences, length of time in the U.S. schools, consistency of schooling, and experience with specific accommodations. This information is compiled, consolidated, and subsequently subjected to a series of decision-making algorithms, which ultimately recommend a set of accommodations most appropriate for each student. The system is designed to adapt to the accommodation policies of different states, and to information collected with different instruments within states or school districts.

The *STELLA* system is explained briefly in Kopriva (2007) and in more detail in Kopriva, Hedgspeth, Koran, and Carr (in preparation).

Does Individual Assignment Make a Difference?

To address this question, a randomized study was completed that examined the effectiveness of assigning accommodations to ELLs with diverse needs and then subsequently assigning the students to one of three groups based on quality of the match between the needs of each of the students and randomly assigned accommodation(s). Primarily, the study asked:

- (1) Is there a difference in the level of performance on a mathematics assessment between ELL students who received the recommended test accommodations, incomplete or not recommended accommodations or no accommodations?

Secondarily, it asked:

- (2) Does sex or type of ELL service significantly interact with accommodation assignments and does this lead to significantly higher test scores for any of these groups?
- (3) If there was a significant difference in performance for the recommended assignment group, as dictated by this study, does the subset of the variables differentially utilized to make those assignments account for the expected portion of the variance in the scores of English language learners?
 - (a) As expected, does ELP-reading (ELP-R) have a significant impact on variance in performance for all students?
 - (b) Since it is expected that ELP-listening (ELP-L) and L1-reading (L1-R) proficiency are differentially effective for different students based on their needs, do these two variables account for a much smaller portion of the variance than ELP-R in overall performance?
- (4) Since it is expected that recommended accommodations will mitigate the effect of English language proficiency issues on test scores, is there a decrease in the correlation between test scores and English language proficiency variables for students who receive recommended accommodations?

Method

Sample

The data utilized in this study were collected as part of the *STELLA* development project conducted in partnership with the State Department of Education in South Carolina. Two counties within the state were identified and targeted on the basis of the large immigrant populations that live in these areas. Within these counties, third and fourth grade ELL teacher specialists who had high proportions of ELL students in their districts were identified; this yielded 11 teachers from five schools. Both third and fourth grade students were selected because researchers and the state agreed that, while the test material was not new to fourth grade students, it should not affect how they interacted with the accommodations they were given.

Through this partnership, 281 Spanish-speaking ELL students were initially recruited for the study. Sufficient data were collected for the $n = 272$ students included in the final analysis. Of these students, 152 were third graders and 120 fourth graders. The students were distributed across low, medium, and high and grade level ELL levels as defined by the participating local educational agencies. The below grade level groups were similar sizes, and the grade level group the smallest ($n_{\text{low}} = 85$, $n_{\text{int}} = 82$, $n_{\text{high}} = 76$, $n_{\text{gl}} = 29$). The students were relatively evenly split by sex with males constituting $n = 146$ or 53.7% of the sample and females constituting $n = 126$ or 46.3% of the sample. Randomization of accommodation assignments occurred evenly across each of the two grades irrespective of the particular type of ELL services provided by participating districts.

Instruments

Teacher form

The teacher specialists who participated in this study were asked to complete a short questionnaire regarding their individual students. Specifically, teachers were asked to provide the name, grade, and type of ELL services the students were receiving. From school records, teachers were asked to rate, on a five-point Likert type scale, the English language proficiency of their ELL students in both reading and listening. This five-point scale was later compressed into a four-point scale by combining the lowest two proficiency levels due to small cell

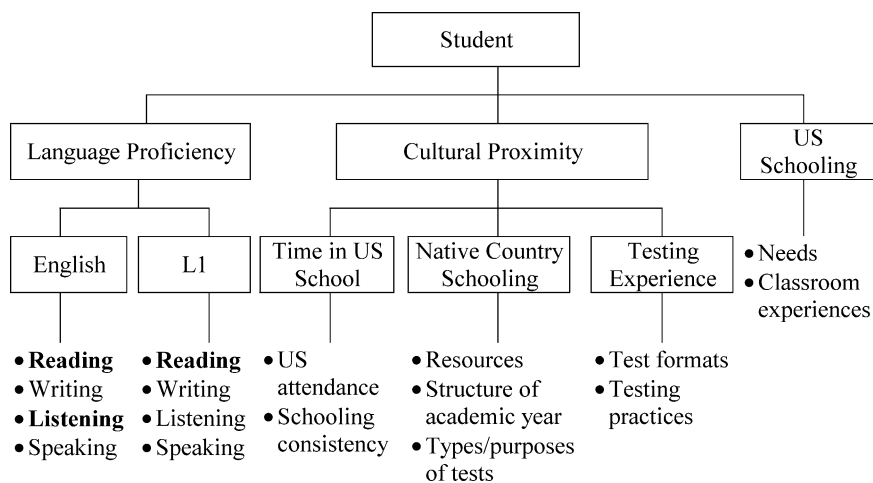


FIGURE 1. *STELLA* decision-making information.

size in the those two groups. Further, from their personal experience, teachers were asked to rate the reading proficiency in their home language of all the students who participated. In some cases, school record information was also available to augment the L1 rating. This information was later used by researchers to determine the appropriateness of the accommodations each student received.

This language proficiency information in the study is a subset of the information used in the *STELLA* decision-making system. Additional student information, including cultural proximity and schooling detail, allows for finer grained accommodation decisions in *STELLA*. For this study, however, only the effectiveness of the first broad cut based on language proficiency was studied. Figure 1 shows the extent of variables possible in the full *STELLA* system, with the study variables bolded under the language proficiency category.

Mathematics assessment

All students who participated in this study were asked to complete a computer-based mathematics assessment. A computer mode was selected in order to facilitate the usage of different test accommodations assigned within each randomized administration. The assessment was developed specifically for this study, was based on South Carolina Department of Education mathematics standards for third grade students, and was judged to be consistent in language and load to local standardized achievement tests. The assessment consisted of 30 multiple-choice and three constructed response items. For the purpose of this study only the

multiple-choice item results were considered.

Procedures

In winter 2004 an administrator from the State Department of Education in South Carolina recruited teacher specialists who provided services for Spanish-speaking students. Once teachers were identified, they were sent the teacher questionnaire to be completed for each of their Spanish-speaking students. These students became the focus of this study.

In April 2005, the participating ELLs completed the English computerized mathematics exam. Students were randomly assigned to either no test accommodations, a picture dictionary, a bilingual glossary, oral reading of test items in English, both oral reading and picture dictionary, both oral reading and bilingual glossary, both picture dictionary and bilingual glossary, or oral reading, bilingual glossary, and picture dictionary. Students who received the accommodation of picture dictionary could click on selected words with their mouse and receive a picture of the word. Students who were given the accommodation of bilingual glossary would receive the Spanish word translation when they clicked on the selected words. Students receiving the oral accommodation automatically had test items read to them in English and could elect to have items repeated. The randomized accommodation package, the number of times students used their provided accommodations, and their scores on the math exam were recorded.

After testing was completed, researchers matched the accommodated

students to their recommended accommodation group based on their ELP-R score, their ELP-L score, and their L1 reading proficiency ratings. While this information is a subset of the data collected in the current *STELLA* system, the randomized assignments followed the general decision-making rules in *STELLA* associated with these variables. After the individualized matching was completed, students who received accommodations were placed in one of two groups, either recommended, or incomplete or not recommended accommodation groups. Those students who received no accommodations remained in the no accommodation group.

Analysis Design

Prior to any analysis of the impact of accommodation group differences, a univariate analysis of variance (ANOVA) was conducted on the eight possible sets of accommodation administrations. This initial analysis was completed to ensure that no accommodation package, in and of itself, resulted in elevated test scores. The randomized assignment of these accommodations across all participating ELL students suggested that no statistically significant differences should be found between the scores of students across all eight administrations conditions and this preliminary analysis was designed to confirm this expectation.

Next, once the students were assigned to the three accommodation groups, an ANOVA analysis was conducted to ascertain if a statistically significant difference existed in test scores for those students who received the recommended accommodations; those who received incomplete or not recommended accommodations and those who received no accommodations. This was followed by a Fisher's LSD pair-wise post hoc comparison of means between the three accommodation groups. It was hypothesized that students in the recommended accommodations group would perform better than the other two groups. While higher performance is not a direct indicator of improved validity of score inferences, lack of access generally decreases student scores for many students. The effective removal or reduction of these barriers to access would then seem to improve the capability of the instrument to measure student ability rather than ancillary abilities such as English language proficiency. This removal or

reduction should tend to lead to increased test scores for those students with some level of content knowledge. As such, significant group differences would indirectly indicate increased validity for the group with the highest test scores (Kopriva, 2000; Abedi, 2001).

After the one-way ANOVA procedure on the accommodation groups was completed, the analysis was expanded to a two-way ANOVA to identify whether student characteristics including grade, sex, and type of ELL service interacted with the results of the accommodation group analysis. It was expected that the results would be consistent across each of these groups, and that while significant main effects were expected for grade and type of ELL service, no significant interactions would be found. Specifically, it was expected that fourth grade students would perform significantly higher than third graders, and that mostly mainstream students would perform significantly higher than students in classes reserved for early ELL students. No significant performance differences were hypothesized between sexes.

Subsequently, a multiple linear regression was performed to determine whether the student variables studied here contributed significantly to student's scores. The student's ELP-R, ELP-L, and L1-R proficiency were selected as predictor variables and test scores were selected as the dependent variable. As the three variables were used in different ways to assign students to the three groups, this analysis

was completed for the entire sample. As per the *STELLA* decision-making rules, it was anticipated that English listening and L1-R would play a secondary role to ELP-R. While it was expected that ELP-R would be important for all students, it is hypothesized that the latter two variables would be differentially effective for different students based on their needs, and that therefore their effectiveness in this summary statistic over students would be substantially smaller.

Finally, correlations of test score results by language proficiency variables were computed for each accommodation group. A decrease in the correlations for the recommended accommodations group would suggest that the impact of these variables on performance was mitigated by appropriate accommodation assignment.

Results

Preliminary Analysis

The 30 multiple-choice items had a reliability index of .684 using Cronbach's alpha. Splitting the data into the three accommodations groups showed that the group with recommended accommodations had the highest reliability, .702, and the incomplete or not recommended accommodations group had the lowest reliability, .640, while the no accommodation group fell in between with .696.

The average test score for the 272 students was 15.540 and the standard deviation was 4.820.

As predicted, the initial ANOVA across the eight possible accommodation packages showed no significant differences in scores, with an $F(7, 264) = .248, p = .973$. This preliminary analysis indicates that no randomized accommodation package had a significant impact on test scores and suggests any differences found in the main analysis would be the results of accommodation assignment. Descriptive data for each of the eight packages are shown in Table 1.

Main Analyses

To determine whether the recommended accommodations group significantly improved ELLs test performance compared to either no accommodation group or incomplete or not recommended accommodations group, a one-way ANOVA was conducted. Table 2 shows the *N*, means, and standard errors for the three accommodation groups. The result of the ANOVA analysis across accommodation groups yielded an $F(2, 269) = 5.989, p = .003$, which indicated that there was a significant difference in test scores between the no accommodation group, the recommended accommodations group, and the incomplete or not recommended accommodation group. A summary of these results can be found in Table 3. Following this analysis, post hoc comparisons were conducted using Fisher's LSD. A significant difference was detected between the no accommodation group and the recommended accommodation group $t(269) = -1.987$,

Table 1. Randomized Accommodation Assignment

	Sex	Grade 3	Grade 4	Total for All Grades	Group N	Mean	SE
No accommodations	F	7	4	11	33	15.000	2.611
	M	11	11	22			
Picture dictionary	F	11	9	20	36	15.556	2.593
	M	8	8	16			
Bilingual glossary	F	10	9	19	36	16.056	2.676
	M	10	7	17			
Oral accommodation	F	12	7	19	33	14.970	2.606
	M	7	7	14			
Oral and picture dictionary	F	2	10	12	29	15.793	2.933
	M	11	6	17			
Oral and bilingual glossary	F	12	4	16	33	15.515	2.701
	M	6	11	17			
Bilingual glossary and picture dictionary	F	11	5	16	37	15.297	2.515
	M	14	7	21			
All accommodations	F	8	5	13	35	16.114	2.724
	M	12	10	22			
Total		152	120	272			

Table 2. Descriptive Statistics for ANOVA

Accommodation Group	N	Mean	SE
No accommodations	33	15.000	.891
Recommended accommodations	94	16.904	.487
Incomplete or not recommended accommodations	145	14.780	.387

Table 3. ANOVA Results for Accommodations Groups

Source	df	MS	F	Significance
Accommodation group	2	134.239	5.989	.003
Error	269	22.413		

$p = .048$, and a significant difference was also detected between the recommended accommodation group and the incomplete or not recommended accommodation group $t(269) = 3.395$, $p < .001$, with the recommended group significantly higher in both cases. Finally, no difference was found between the no accommodation group and the incomplete or not recommended accommodation group $t(269) = 0.241$, $p = .809$. A summary of the post hoc results can be found in Table 4.¹

Subsequently, because of sample size, three separate two-way ANOVAs were run to test the interaction effects of accommodation group by sex, grade and type of ELL services. Group sizes and descriptive statistics for these three student descriptor variables are shown in Table 5. Although all analysis resulted in significant group differences ($F(2,269) = 3.304$, $p = .007$ for the analysis which included the sex variable, $F(2,269) = 6.183$, $p < .001$ for grade, $F(2,269) = 12.907$, $p <$

.001 for type of service), the group differences were for the main effects only. As hypothesized, for all three two-way ANOVAs, no interaction effects were found, which indicates there was no evidence of significant interaction between accommodation group and sex, between accommodation group and grade, and between accommodation group and type of service, as shown in Table 6.

The main effect differences varied as expected: by grade (grade 4 students scoring on average 2.388 points higher than grade 3, $p < .001$); by ELL service (mainstreamed students on average scored 4.239 points higher than students in limited English proficiency classes, $p < .001$) and no significant difference between male and females.

Correlations and Regression Analyses

Correlations were computed to determine the relationship between test scores and language proficiency vari-

ables for the entire sample and a regression was performed to investigate how well the language proficiency variables predicted test scores for all students. Subsequently, correlations were computed in order to address the hypothesis that the relationship between test scores and each of the language proficiency variables would decrease in the recommended accommodations group as compared to the other two groups.

As Table 7 indicates, for the total sample all correlations were significantly different from zero unless otherwise noted. Table 8 illustrates that the regression analysis of data from the total sample yielded an $R^2 = .281$, $p < .001$, and ELP-R was found to contribute to the regression at a slope significantly different from zero ($b = 2.669$, $p < .001$). This finding suggests ELP-R is a salient predictor of test score results and that students with higher test scores are associated with higher ELP reading proficiency. Further, the impact of LI-R was also found to be significantly correlated with total test score, though, as noted by the size of the regression weight and its significance level, its impact was more muted than that of ELP-R. Contrary to expectations in this study, the impact of ELP-L was not found to be significantly different from zero. However, the very high correlation between the two variables ($r = .9$) suggests that the data for reading and listening used in this study explained similar portions of the variance in student score.

Table 9 shows that the correlation between test score and the English language proficiency variable for reading did in fact decrease for the group of

Table 4. Fisher LSD Results for Accommodations Groups

Group 1	Group 2	Mean Difference (1-2)	SE	Effect Size	Significance
No accommodations	Recommended accommodations	-1.904	.958	-1.987	.048
	Incomplete or not recommended accommodations	.221	.913	.242	.809
Recommended accommodations	No accommodations	1.904	.958	1.987	.048
	Incomplete or not recommended accommodations	2.125	.626	3.395	<.001
Incomplete or not recommended accommodations	No accommodations				
	Recommended accommodations				

Table 5. Student Test Scores for Two-Way ANOVA

Group	N	Mean	SE
Sex			
F	126	15.6984	1.399
M	146	15.4041	1.275
Grade			
3.0	152	14.4868	1.175
4.0	120	16.8750	1.540
Type of ELL service			
Full	175	14.0286	1.060
Mainstream	97	18.2680	1.855

Table 6. Interaction Results, Two-Way ANOVA

Group Interaction	MS	F	Significance
Accommodation group and Sex	46.954	2.106	.124
Accommodation group and Grade	5.252	.248	.781
Accommodation group and Type of service	.706	.037	.964

students, with recommended accommodations ($r = .420$) compared to both the no accommodation group ($r = .539$) and the incomplete or not recommended accommodation group ($r = .549$). The correlation between test score and English language proficiency for listening variable follows the same general pattern. However it is interesting that the amount of decrease in ELP-L is reasonably different from the ELP-R, and is the most pronounced when no accommodations are received. The correlation between test score and L1-R was less in the recommended accommodations group than for students who received no accommodations but, contrary to expectations, correlation between test scores and L1-R was virtually nonexistent in the incomplete and not recommended accommodations group. Unlike the correlations for the ELP variables, the L1-R relationships were not significantly different from zero.

Table 7. Total Sample Correlations

Variable	Total Score	ELP-R	ELP-L	L1-R Proficiency
Total score	1.000	.511	.431	.112
ELP-R		1.000	.900	-.027 ns
ELP-L			1.000	.046 ns
L1-R				1.000

to significantly mask the potential benefits of particular test accommodations for students who need them.

The findings suggest the relevance of the language proficiency variables and their relationship to student test scores, especially English language proficiency in reading. It is worth remembering that, while the L1 Reading effect was muted in the regression results for the overall sample and not significant in the correlations by group, this may be reasonable given the limitations of the study. Teacher report of this variable was the weakest measure in the study, the distribution was rather flat, and so the results were not surprising. Further, the study used a test in English and the L1 effect here is limited to translation of selected nouns in the bilingual glossary accommodation, an accommodation provided to L1 literate students only. In a related study, researchers found that lack of L1 literacy was also a factor that seems to influence selection (Kopriva, Winter, Wiley, Emick, & Chen, in press), but the accommodations for this option were not studied here. It may be, as ELL experts (e.g., August, Calderón, & Gottlieb, 2004) suggest, that this factor is important but that the whole spectrum of the variable needs to be properly accommodated for the effect to be properly observed. Finally, although this study suggests that ELP reading and listening are measuring almost the same constructs, analyses of English language proficiency tests indicate unique variance in each of these two variables (for instance see Lara, Winter, Kopriva, Ferrara, & Bunch, 2007). Future studies will need to determine how important ELP listening is to the overall selection process. Both the L1-R and ELP-L results from this study have been noted in *STELLA*, where a triangulation of data from other sources (tests and parents for evidence of L1-R, and tests and teachers for evidence of English listening skills) are now being used in the assignment algorithms.

One of the key, albeit indirect, findings in this study is that when the impact of individual accommodations or packages on test scores is analyzed without regard to student need (ANOVA results under *Preliminary Analyses* and Table 1), the differences are not significant. This is not only important in light of the later ANOVA results that indicate significant differences when matching

Table 8. Total Sample Linear Regression Results of Language Proficiency Variables on Test Scores

Independent Variable	<i>b</i>	<i>SE</i>	Significance
Constant	7.585	1.053	<.001
English language proficiency in reading	2.669	.494	<.001
English language proficiency in listening	-.612	.510	.231
L1-R proficiency	1.435	.606	.019

Table 9. Group Correlations between Test Score and Language Proficiency Variables

Group	<i>N</i>	ELP-R	ELP-L	L1-R
No accommodations	33	.539	.555	.235 (ns)
Recommended accommodations	94	.420	.372	.132 (ns)
Incomplete or not recommended accommodations	145	.549	.430	.000 (ns)

of need to accommodation is undertaken. The findings are also important because this is how research on accommodations has typically been carried out, by package and using a broad group of ELLs who are not screened by need. The differences between the two analyses highlight the danger of misinterpreting past research results and a primary reason why findings over studies may not be consistent.

It will be important to replicate this study with the full *STELLA* system, which includes overlapping data from three sources. Koran et al. (2006) found that the other sources provide both triangulated support as well as unique information about several of the components, and that, while the composite tended to be consistent across sources, additional data could potentially be useful to fine tune decision-making practices beyond what are reported here. Because of the randomized design of the current study, however, there seems to be every reason to expect that the findings described in this paper provide a reasonably accurate reflection of improved validity inferences when students are properly accommodated.

To date, the differential effects of other variables used in *STELLA* to make finer decisions regarding recommended accommodations, such as cultural proximity and U.S. schooling experiences, have not been conducted. Future studies should also be conducted to examine the effects of use of accommodations in classrooms, and to examine if these findings hold for a

more robust sample of ELLs. Limitations of the current study are that the vast majority of students represented in this sample were of Latina/o ethnic backgrounds and were elementary students only. Research with students from differing language, ethnic, and national origins should be considered as well as studies that examine the differential assignment based on age and possibly content areas.

Overall, the study findings suggest that a more methodical approach toward assigning accommodations to ELL students would be beneficial compared to current practices. Additional research using systematized decision-making processes are necessary to validate the findings reported here, and to reconsider the effectiveness of many of the accommodations that other work has suggested might or might not be useful for this population.

Conclusion

Recommended accommodations appear to be critical for the better assessment of special populations in academic accountability programs. Our results suggest that at the individual level, when accommodation decisions are not appropriate to meet the need of the student, test results can misrepresent what the student knows and can do. At the test aggregate level, consistent and recommended accommodation decision making is critical to the validity of score comparisons across schools, programs, districts, and populations of students (Fuchs et al., 2000;

Hollenbeck, Tindal, & Almond, 1998; Kopriva, 2000). There has been a strong call for more systematic methods associated with selecting appropriate large-scale test accommodations for students in special populations (e.g., Abedi et al., 2003; Kopriva & Mislevy, 2001; Thurlow, Elliott, & Ysseldyke, 2003). Our results provide empirical support for the need to reexamine the current methods of assigning accommodations and suggest that individual characteristics need to be considered. Additionally, the study shows evidence of support for the use of the *STELLA* system as one way to address the lack of systematicity in assigning accommodations.

Note

¹The assumption of equality of variance was found to hold for this analysis $F(2, 269) = 1.032, p = .358$.

References

- Abedi, J. (2001). *Assessment and accommodation for English language learners: Issues and recommendations*. CRESST Policy Brief 4. Los Angeles: University of California, Center for the Study of Evaluation.
- Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners in NAEP* (CSE Tech. Rep. No. 586). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Adelman, L. (1992). *Evaluating decision support and expert systems*. New York: Wiley.
- Albus, D., Thurlow, M., Liu, K., & Bielinski, J. (2005). Reading test performance of English-language learners using an English dictionary. *The Journal of Educational Research, 98* (4), 245–253.
- August, D., Calderón, M., & Gottlieb, M. (2004). *How to design science-based instructional systems for achieving academic English and content*. Paper presented at OELA Celebrate Our Rising Stars Third Annual Summit, Washington, DC.
- Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations*. CSE Technical Report 448. Los Angeles: University of California, Center for the Study of Evaluation.
- Butler, F. A., & Stevens, R. (2001). Standardized assessment of content knowledge of English language learners K-12: Current trends and old dilemmas. *Language Testing, 18* (4), 409–427.
- Castellon-Wellington, M. (2000). *The impact of preference for accommodations: The performance of English language*

- learners on large-scale academic achievement tests. CSE Technical Report 524. Los Angeles: University of California, Center for the Study of Evaluation.
- Douglas, K. (2005). *Qualitative analysis of TTELL focus group data*. Unpublished, manuscript, University of Maryland College Park.
- Education Law Center (2002). Complaint to the Office for Civil Rights Concerning Pennsylvania's Testing of English Language Learners (<http://www.edlawcenter.org/>) (accessed March, 2004).
- Emick, J., Kopriva, R., Chen, C., Mislevy, R., & Carr, T. (2006). *Achieving accurate results for diverse learners: Access-based item development*. Paper presented at the American Educational Research Association National Meeting, San Francisco, CA.
- Emick, J., Monroe, R., Kopriva, R., & Sprehn, M. (2007). *The culture of the U. S. testing system: A novel response*. Submitted for publication.
- Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. B., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, 67(1), 67–81.
- Hafner, A. L. (2001). *Evaluating the impact of test accommodations on test scores of LEP students and Non-LEP students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.
- Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education*, 16 (2), 159–188.
- Hollenbeck, K., Tindal, G., & Almond P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *Journal of Special Education*, 32(3), 175–183.
- Improving America's School Act, Pub. L. No. 103–382 (1994).
- Koenig, J. A., & Bachman, L. F. (Eds.) (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessment*. Washington, DC: National Research Council, National Academies Press.
- Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.
- Kopriva, R. (2002). *Taxonomy for testing English language learners*. Funded proposal to the U.S. Department of Education, Grants for Enhanced Assessment Instruments Program.
- Kopriva, R. (2005). *Introducing the STELLA system*. Presentation at the meeting of the National Conference on Large-Scale Assessment, San Antonio, TX.
- Kopriva, R. J. (2007). *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- Kopriva, R. J., & Hedgspeth, C. (2005). *Technical manual, selection taxonomy for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- Kopriva, R. J., Hedgspeth, C., Koran, J., & Carr, T. (in preparation). *Introducing STELLA: An individualized testing accommodation assignment tool*.
- Kopriva, R., & Koran, J. (2006). *Validity of a matching system for assigning accommodations for individual English language learners*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Kopriva, R. J., Koran, J., & Hedgspeth, C. (2007). The importance of systematically matching student needs and test accommodations. In L. Cook & C. Cahan (Eds.), *Large scale assessment and accommodations: What works?* Washington, DC: Council of Exceptional Children Press.
- Kopriva, R., & Mislevy, R. J. (2001). *Valid assessment of English language learners*. Funded proposal to the U.S. Department of Education, Institute of Education Sciences, Washington, DC.
- Kopriva, R. J., Winter, P., Wiley, D.E., Emick, J., & Chen, C. S. (in press). Estimating the impact of altering mathematics items to address access. *Educational Assessment*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- Koran, J., & Kopriva, R. J. (2006, April). *Teacher assignment of test accommodations for English language learners*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Koran, J., Kopriva, R., Emick, J., Monroe, R., & Garavaglia, D. (2006). *Teacher assignment and a multi-source computerized approach for making individualized test accommodation decisions for English language learners*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Lara, J., & August, D. (1996). *Systemic reform and limited English proficient students*. Washington, DC: Council of Chief State School Officers.
- Lara, J., Winter, P., Kopriva, R., Ferrara, S., & Bunch, M. (2007). English language development assessment. In J. Abedi (Ed.), *English language proficiency tests: Past, present and future*. Davis, CA: U.C. Davis Press.
- Liu, K. K., Anderson, M.E., Swierzbis, B., Spicuzza, R., & Thurlow, M. L. (1999). *Feasibility and practicality of a decision making tool for standards testing of students with limited English proficiency* (Minnesota Report 22). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Mann, H., Emick, J., Cho, M., & Kopriva, R. (2006). *Addressing the validity of test score inferences for English language learners with limited proficiency using language liaisons and other accommodations*. Presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.
- No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2002).
- Plake, B., & Impara, J. (2006). *Accommodation station expert panel meeting*. Savannah, GA.
- Rivera, C., & Collum, E. (Eds.) (2006). *A national review of state assessment policy and practice for English language learners*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Rivera, C., & Stansfield, C. W. (2003). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment*, 9 (3/4), 79–105.
- Samuelsen, K., & Kopriva, R. J. (2004). *Making sure students receive recommended accommodations on academic tests*. Presentation at the National Summit Sponsored by the Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students, Washington, DC.
- Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements* (2nd ed.). Thousand Oaks, CA: Corwin Press.