2017

Framing Appropriate Accommodations in Terms of Individual Need: Examining the Fit of Four Approaches to Selecting Test Accommodations of English Language Learners

Jennifer Koran
Southern Illinois University Carbondale, jennifer.koran@gmail.com

Rebecca J. Kopriva
University of Wisconsin

This is an Accepted Manuscript of an article published by Taylor & Francis in *Applied Measurement in Education*, available online: http://www.tandfonline.com/doi/full/10.1080/08957347.2016.1243539

FIT OF ACCOMMODATIONS

Framing Appropriate Accommodations in Terms of Individual Need: Examining the Fit of Four Approaches to Selecting Test Accommodations of English Language Learners

Jennifer Koran

Southern Illinois University Carbondale

Rebecca J. Kopriva

University of Wisconsin

Author Note

Jennifer Koran, Quantitative Methods Program, Southern Illinois University Carbondale; Rebecca J. Kopriva, Wisconsin Center for Educational Research, University of Wisconsin.

This research was supported by a Title I Enhanced Assessment Grant Application Number S368A030005. The contents of this article reflect the views of the researchers and do not represent the policy of the Department of Education, nor is endorsement by the federal government implied. The authors wish to acknowledge the following persons for their assistance with this research: Jessica Emick, Diane Garavaglia, Carlos Hipolito-Delgado, Kirsten Lennon, Heather Mann, J. Ryan Monroe, Karen Samuelsen, and Daisy Wise.

Correspondence concerning this article should be addressed to Jennifer Koran, Quantitative Methods Program, Wham 223 MC 4618, Southern Illinois University Carbondale, 625 Wham Drive, Carbondale, IL 62901. E-mail: jkoran@siu.edu

Abstract

Providing appropriate test accommodations to most English language learners (ELLs) is important to facilitate meaningful inferences about learning. This study compared teacher large-scale test accommodation recommendations to those from a literature- and practitioner-grounded accommodation selection taxonomy. The taxonomy links student-specific needs, strengths and schooling experiences to large-scale test accommodation recommendations that differentially minimize barriers of access for students with different profiles. A blind panel of experts rated four sets of recommendations for each of 114 ELLs. Results found the taxonomy was a significantly better fit for distinguishing accommodations by student need than teacher recommendations. Further, the fit of teacher recommendations showed no difference when the teacher used a structured data collection procedure to gather profile information about each of their ELLs and when they did not, and teachers' recommendations were not found to differ significantly from a random set of accommodations. Findings are consistent with previous literature that suggests the task of matching specific accommodations to individual needs, rather than the task of identifying individual needs, is where teachers struggle in recommending appropriate test accommodations.

Framing Appropriate Accommodations in Terms of Individual Need: Examining the Fit of Four Approaches to Selecting Test Accommodations of English Language Learners

The assignment of accommodations that would effectively ameliorate barriers to traditional testing procedures for individual students who need them is critical for the valid large-scale content assessment of special populations in academic accountability programs. Recognizing the urgency of this need, experts have made a strong call for more systematic methods associated with selecting appropriate test accommodations for students in special populations (see Thurlow & Kopriva, 2015, for a review). Improvement in assigning accommodations for large-scale tests like those used by states is especially critical for English language learners (ELLs), a tremendously diverse group that has a relatively short history of inclusion in these assessments.

While lagging behind accommodations research for other special populations, such as students with disabilities, the field has recently begun to focus on developing systems for selecting appropriate test accommodations for ELLs. Current practice in selecting accommodations for ELLs, however, typically consists of an unstructured process that relies heavily on the judgment of each individual student's teacher (Kopriva & Koran, 2008). Nascent research both on developing accommodation selection systems and the effectiveness of current practice in selecting test accommodations for ELLs is scarce. This study provides a much-needed comparison between an accommodation selection system for ELLs and current practice.

Literature Review

Appropriate Test Accommodations for ELLs

Appropriate large-scale test accommodations are important for accurate inferences at multiple levels in inclusive state and national accountability programs. At the individual level when accommodation decisions are not appropriate to meet the need of the student, test results often misrepresent what the student knows and can do (Kopriva, Thurlow, Perie, Lazarus & Clark, in press). Thus, providing appropriate test accommodations can result in more meaningful inferences about an individual student's abilities for parents and teachers. At the aggregate level, consistent and appropriate accommodation decision-making is critical to the validity of program comparisons and large-scale test score comparisons across states, districts and schools (Kopriva & Lara, 2009). Greater consistency in providing appropriate accommodations can result in more meaningful inferences about the learning taking place in different classrooms, schools, and districts for both school administrators and state education agencies.

While the consequences are important, the complexity of selecting accommodations is likewise demanding. It is not simply a matter of providing test accommodations; those accommodations must be appropriately matched to the needs of the individual student. One study showed that ELLs who received inappropriate accommodations performed at a level comparable to ELLs who received no accommodations on a mathematics assessment; students receiving appropriate accommodations significantly outperformed both of these other two groups (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007). It involves more than simply identifying the student's needs. Complex interactions of factors must be taken into account: the individual's English language proficiency, first language literacy, the language of instruction, and skill in using bilingual dictionaries and translation aids, to name a few (Pennock-Roman & Rivera, 2011; Solano-Flores, 2014). Further, ELLs often have compensatory strengths that they can draw upon to demonstrate their knowledge on tests, if given the opportunity (Del Rosario

Basterra, Trumbull, & Solano-Flores, 2011). Understanding the test-related constellation of factors associated with each ELL is the first step. This constellation then must be mapped to an appropriate set of test accommodations, which will help alleviate the student's need while not giving an unfair advantage or overwhelming the student.

Research on Systematic Methods for Selecting Test Accommodations for ELLs

Kopriva and Koran (2008) provide a review of systems for selecting appropriate large-scale test accommodations for students in special populations. One approach to making more systematic accommodation decisions for students with disabilities has used an inductive approach, systematically testing each student with different accommodations and examining which accommodations give the student a differential boost in test performance (Fuchs, Fuchs, Eaton, & Hamlett, 2005; Fuchs, et al., 2000). This approach has a strong empirical grounding but can be extremely time consuming and may not be suitable for selecting accommodations in large-scale content assessment settings. In the last 20 years researchers have been working on developing deductive methods for matching individual students with relevant test accommodations. These deductive methods rely on building accommodation selection guidance based on theories of the interaction between relevant student characteristics (strengths and weaknesses) and test accommodations. An important prerequisite for the systematic application of theory in this context is to collect accurate and relevant data about the student to use as the basis for decision making. Some applications of the deductive approach rely on direct assessment of student characteristics to collect this data (e.g. the Accommodation Station; Ketterlin-Geller, 2003; Tindal, 2006). Others rely on the structured reports of informants, such as parents and teachers, and the collection of extant data about the student, such as English language proficiency test scores (e.g. Kopriva, Carr, & Cho, 2006). The current large-scale test

consortia, Smarter Balanced and PARCC, are both interested in building student profile systems in order to properly assign accommodations to ELLs and students with disabilities (Thurlow & Kopriva, 2015).

Regardless of the approach taken in collecting relevant information about the student, what the deductive methods have in common is that a theory-based decision tree is applied to the student information to select accommodations that are a most appropriate match to that student's particular characteristics. The development of these rubrics to make the best use of complex information in diverse populations is a broad undertaking. Thus, while some research has indicated great potential for these systems, other studies have failed to show the intended effects. Research on test accommodation selection systems using the deductive approach is ongoing.

Research on Teacher Assignment of Test Accommodations

Current practice in selecting accommodations for ELLs typically consists of an unstructured process (Albus, Thurlow, Liu, & Bielinski, 2005; Rivera & Collum, 2006) that relies heavily on the judgment of each individual student's language and/or content teacher (Council of Chief State School Officers, 2003; Liu, et al., 1999; Thurlow & Kopriva, 2015). Very little research has examined the quality of teacher judgment in selecting large-scale test accommodations for ELLs. However, some studies that have examined the quality of teacher judgment in selecting test accommodations for students with disabilities have found that teachers predict at little above chance level which students will benefit from which accommodations (Fuchs, et al., 2000; Helwig & Tindal, 2003, Weston, 2003; Plake & Impara, 2006).

Purpose of the Current Study

The purpose of the present study is to see if there are differences in how well the needs of ELLs are accommodated on large-scale standardized content assessments as recommended by

different approaches. Two sets of teacher-recommended accommodations, accommodations generated from a research-based individuated accommodation taxonomy system, and one random set of accommodations were compared.

Methods

Nineteen teachers from different grades collected relevant information about ELLs participating in the study, and four sets of recommended accommodations were generated for each ELL. A panel of ELL experts was subsequently convened to rate the appropriateness of the four sets of accommodations in meeting the needs of each student and a linear mixed-effects model was fitted to the rating data to compare the approaches.

Participants

Teachers of ELLs from the states of Maryland (four teachers), North Carolina (four teachers), and Texas (11 teachers) agreed to participate in the study. Eight of the teachers had been teaching ELLs for 1-5 years, five of the teachers had been teaching ELLs for 6-10 years, and six of the teachers had been teaching ELLs for 11 or more years. Each teacher recruited parents of six ELLs who represented a range of English skill levels in their classes, for a total of 114 parent-student dyads in grades K-12. The parents agreed to participate in an interview/interpreter with the teacher and to allow the teacher to use his/her child's information in the study.

The students spoke a variety of languages in the home including Spanish (57.9%), Vietnamese (14.9%), Chinese (5.3%), Arabic (3.5%), and other home languages (18.5%). There was at least one student in each grade level, but most (68.4 percent) of the students were in grades three through nine. Students were enrolled in a variety of language programs, such as bilingual programs (9.7 percent), self-contained ELL classroom programs (32.5 percent), and

pull-out ELL programs (30.7 percent), with 27.2 percent of students in some other type of program, such as programs that use a combination of these approaches. The students were roughly evenly distributed across four broad levels of English language proficiency ranging from students who had few or no English skills to students whose skills in English allowed them to keep pace with their monolingual English-speaking peers in their content classrooms. Likewise, the students' parents gave ratings of native language proficiency (again in reading, writing, listening, and speaking) that suggested that this sample of students was roughly evenly distributed across three levels of native language proficiency (low, medium and high).

Instruments

The major instruments used in this study were two teacher open-ended recommendation surveys to be completed at different stages in the study, and the parent, teacher, and records questionnaires associated with the STELLA recommendation system (the Selection Taxonomy for English Language Learner Accommodations, Kopriva, Carr & Cho, 2006), an empirically-based system for selecting test accommodations for ELLs in grades 3-12.

Teacher recommendations. Two teacher recommendation surveys (available in supplemental files from the authors upon request) were created for teachers to recommend a specific set of test accommodations for each of their six students participating in the study. Each survey consisted of one question. The first teacher accommodation survey asked teachers to select accommodations for their students based on their current local test accommodations procedure and was completed before teachers began completing the parent, teacher, and records structured data collection questionnaires that were created to collect data for STELLA taxonomy. The second teacher accommodation survey was completed after teachers had finished the data collection protocol. This survey asked the teachers to select accommodations for each of their

students on the basis of the student information they had collected in the questionnaires. The accommodations recommendations made by the teachers on the questionnaires had no influence on recommendations made by the STELLA system.

STELLA. The STELLA system collects data about student strengths and challenges and links it to promising accommodations for students with different profiles. These student data, collected from parents, teachers and records, include proficiency levels in English and their native language across the four domains of reading, writing, listening and speaking, consistency and structure of their schooling, and their classroom experiences such as types and methods of student evaluations in their home countries (as relevant) and in their US schools. The accommodation options in STELLA are those identified in meta-studies such as Pennock-Roman and Rivera (2011) that research and practice have suggested are promising for ELLs. The decision-making taxonomy consolidates the data into student-specific profiles and then uses theory- and expert judgment-based decision trees to systematically match student profiles to recommended large-scale accommodations. A discussion of the specific types of student information STELLA collects and how the individual student profiles are used to arrive at the test accommodation recommendations can be found in Kopriva and Myers (2016). While the focus of this work is to specify useful on-demand large-scale content test accommodations, educator feedback has suggested that teachers found the methods used in STELLA to be helpful in accommodating their ELLs in their content classrooms (Kopriva, Carr & Cho, 2006).

The working prototypes (beta version) of the student data questionnaires and the decision trees were used in this study. Federally-funded development of the STELLA questionnaires and the consolidation and decision-making algorithms included a nationwide review of test accommodation policies and the extant literature, teacher focus groups, parent and teacher

interviews, ongoing external reviews and oversight of products by state development partners, and two panels who reviewed the taxonomy decision trees specifically, a panel of state ELL educators from around the country and an expert panel (Douglas, 2005; Kopriva et al., 2006; Kopriva & Koran, 2008). More detailed explanations of the student data collected by the system and the qualitative findings during development may be found in Kopriva, Koran, and Hedgspeth, (2007) and a related experimental study (Kopriva et al., 2007). A white paper discusses the decision algorithms in detail (Myers & Kopriva, 2015).

Data collection questionnaires. The STELLA data collection surveys consist of three forms that systematically structure the collection of information about an individual student that is relevant to understanding the student's need for standardized test accommodations and that student's strengths and experiences relevant to making use of accommodations. Table 1 provides a brief listing of the information collected through each of the three forms in the STELLA data collection protocol. The three different forms correspond to three different sources of relevant information: the student's school record, parent, and teacher. The Record Form collects information that is in the student's file at the school and for this study it was completed by the student's teacher. The Parent/Guardian Interview Form is a questionnaire with an interview protocol that is facilitated in this study by the participating teacher (with the aid of an interpreter, if necessary). The Teacher Form collects observations the teacher has made about the student based on classroom experience. The data collection questionnaires are designed so that some information is duplicated across forms (e.g. parent ratings of native language proficiency and teacher ratings of native language proficiency) for the purposes of triangulation. More detailed information about the three forms is available from the authors upon request.

Insert Table 1 about here

Procedures

Data collections. After teachers and students/parents were identified and agreed to participate, the first accommodation survey was completed by the teachers and served as an ecologically valid baseline against which to compare later large-scale test accommodations recommendations (teacher before). Next, the teachers completed the STELLA parent interview and teacher and records data collections. It could be argued that the STELLA questionnaires provided targeted information that might give teachers greater insight into their students. Thus, the teachers were next asked to complete the second teacher test accommodation survey (teacher after). The function of the second set of teacher recommended accommodations was to "level the playing field" with the STELLA system by assuring that the teacher had access to the same information about the student.

Next, the STELLA decision taxonomy was applied to the student data from the questionnaires to produce a third set of test accommodations recommended for each student (STELLA). Finally, a random set of accommodations was drawn for each student from all sets of accommodations from the first three sources for all other students in the study (random). Sets of accommodations that had been proposed for real students were used so as to avoid random sets with implausible combinations of accommodations.

Ratings. Four ELL experts were convened to form an independent evaluation panel. Three of the panelists were teachers from three different districts within the state of Maryland and had classroom experience with ELL assessment and accommodations as well as masters degrees in education with specializations in ESOL/bilingual education and multicultural teacher education. The remaining panelist was a researcher who had previous experience as a classroom

teacher and also experience with ELL testing issues in related test accommodations research. These four panelists were independent from the teachers who had completed the data collection protocol and recommended accommodations for the students. One of the authors provided rating panel members with additional training in test accommodations, as described later in this section.

The materials that the raters viewed were carefully prepared to maintain the confidentiality of the participants and minimize systematic bias in the ratings. Student and teacher names were removed from the STELLA questionnaire forms, the four different sets of accommodation recommendations for each student were presented in a common format to mask the source of the recommendation, and the four sets of accommodations were randomly ordered for each student and were labeled according to their random order.

Raters were trained to examine the student information found in the questionnaire forms, and rate each of several proposed sets of accommodations for its appropriateness in meeting the individual student's test accommodation need. Rater training included orientation to the three forms, and definitions of the specific accommodations in the proposed sets of accommodations. Raters were introduced to a seven-point holistic rating scale to answer the question "How optimal is this set of accommodations for this student?" The scale ranged from completely optimal (1) to completely inappropriate (7). To complete the training raters were given materials for three fabricated students along with three sets of proposed accommodations for each student. The raters used the scale to rate the accommodations and discussed their ratings and their reasoning until they reached consensus.

Analysis

To answer the question of whether there are systematic differences in the appropriateness, or fit, of the recommended sets of accommodations relative to the student data

collected about profiles of needs, challenges and contextualized demographic information, a linear mixed effect model was fitted to the ratings. This technique effectively accommodates hierarchically structured data, repeated measures, and missing observations. The data have a hierarchical structure with students nested within teachers. This is an important consideration because students with the same teacher will tend to have more in common with one another than students with different teachers. It also means that teacher serves as a hierarchical data structure in the analysis as well as a source of accommodation recommendation. The ratings associated with multiple raters represent repeated measurements on each set of accommodations. Finally, the linear mixed effects model easily accounts for ratings missing completely at random, as one rater did not have enough time to complete ratings for all students in the study. We have elected to describe the model in words because this approach happens to be simpler and more succinct for this particular model.

Rating was the dependent variable. In order to address the main purpose of this study, source of the large-scale accommodation recommendation (*Source*; four levels) was included as fixed effect in the model. *Source* was structured as three dummy codes, contrasting teacher-before, teacher-after, and STELLA sources with the random source as the baseline; in this context teacher (teacher-before and teacher-after) is two of the sources of accommodation recommendation. In addition, the effect of the leniency of the rater was controlled by including rater as a random effect by teacher in the model; in this context teacher is a hierarchical data structure within the model. Again, *Rater* was structured as three dummy codes, contrasting the first three raters with the fourth rater. *Rater* was also treated as a repeated measure, thus allowing a separate residual variance to be estimated for each rater. To support the tenability of the main conclusion, the *Source*Rater* interaction was also included in the model and assessed for

statistical significance. This assessed whether there was any evidence of systematic bias due one or more of the raters showing systematic partiality toward any of the sources of sets of accommodations. The intercept was also included in the model and treated as a random effect by student. Maximum likelihood estimation in SAS Proc MIXED was used to estimate the model parameters.

Results

There were four sets of accommodations recommended for each of the 114 students. Thus, there were 456 cases in all (*Source* by student combinations) for each rater to review. One rater did not rate all of the cases, so there were 124 cases (27.2%) that only had three ratings. The remaining 332 accommodation-student combinations had complete data (four ratings).

Table 2 displays descriptive statistics for ratings associated with accommodations recommendations from the four sources. STELLA accommodations recommendations had the lowest mean, indicating the most appropriate fit to the students' needs on average. The minimum and maximum values indicate that the full range of the rating scale was used with all accommodation sources. Eighty-nine percent of the total variability in the ratings is among the repeated measures within each student. Six percent of the total variability in the ratings is across students and five percent of the total variability is across teachers (teacher as data structure). These latter two percentages suggest that the multilevel analysis accounting for the nested structure of the data is appropriate.

Insert Table 2 about here

The *Source*Rater* interaction was not significant, $F(9,1517)=1.33$, $p=0.22$ indicating that there is no evidence to suggest that individual raters were differentially partial to particular

sources of accommodation recommendations. Thus, the main effects of *Rater* and *Source* can be generalized across sources and raters, respectively. The effect of *Rater* was significant, $F(3,54)=43.25$, $p<0.0001$ suggesting it is appropriate to keep this term in the model to control for the differing effects of the relative harshness or leniency of different raters. The effect of *Source* was also significant, $F(3,1517)=116.14$, $p<0.0001$. This finding supports the idea that the ratings differed systematically depending on the source of the accommodation recommendation.

Additional tests illuminate the nature of the effect of *Source*. The ratings associated with the STELLA accommodations were significantly different from ratings associated with the randomly assigned accommodations, $t(1517)=-10.24$, $p<0.0001$. Tukey-Kramer post-hoc tests showed that the ratings given to the STELLA recommendations were significantly different from the two sets of teacher recommendations, teacher-before: $t(1517)=-14.94$, $p<.0001$; teacher-after: $t(1517)=-15.33$, $p<.0001$ (teacher as accommodation source). The ratings associated with teacher-before, $t(1517)=-0.68$, $p=0.50$, and teacher-after, $t(1517)=0.00$, *n.s.*, recommendations were not significantly different from random (teacher as accommodation source). Post-hoc tests also showed no statistically significant differences between the ratings given to teacher (before) and teacher (after) recommendations, $t(1517)=-0.39$, $p=.9797$ (teacher as accommodation source).

The results show small but statistically significant variation associated with both students, $z=1.96$, $p=0.03$, and teachers, $z=4.30$, $p<0.001$ (teacher as data structure). While the variation among rater effects that can be attributed to differences among teachers (teacher as data structure) is statistically significant, $z=2.35$, $p=0.01$, its magnitude is about half of the value of the remaining systematic variance attributable to teachers (teacher as data structure). Error variance attributable to individual raters is more substantial, ranging from 1.15 (rater 4) to 1.75

(rater 1). Square roots of these values suggest that the variability in the ratings unexplained by the model amounts to a standard deviation of a little over one scale point on the seven-point scale. Thus, we may consider the standard error of measurement of the rating scale to be approximately one scale point and varying somewhat across raters.

Discussion

The analysis of the quality of the recommended accommodations, as evaluated by the expert raters, indicates that accommodations recommended by the STELLA system are rated as providing a significantly better, and we argue more appropriate, fit between characteristics of individual students and accommodations than did the accommodations recommended by the teachers. In addition, teachers' recommendations before and after completing the structured data collection procedure were not significantly different from each other or from a random set of accommodations which had been recommended for a different student.

There are several common challenges in designing test accommodations research studies: the heterogeneity of the population, the breadth of options to be considered, and the considerable demand of educational assessment. Most every study in ELL test accommodations research compromises in at least one of these areas for the sake of feasibility. Some studies narrow the population. For example, a study may only consider ELLs in a particular bilingual education program. Thus, the population may have L1 and cultural commonalities, in addition to perhaps being in the same grade level, limiting the generalizability of the findings. Other studies may limit the breadth of options considered. For example, the Kopriva, et al. (2007) experiment considered only a few popular accommodations packages. Finally, some studies may maintain the heterogeneity of the ELL population and the breadth of accommodation options considered, but then must sacrifice the testing of the accommodations in classroom assessments with ELLs.

This study did not research accommodations per se, but did research the abilities of four sources of recommendations to fit data linked to individual student needs and strengths. By linking specific accommodations to student characteristics associated directly with the measurement of content concepts and skills, developers argue that the STELLA recommended accommodations would be able to better minimize key barriers to traditional testing procedures for students with particular profiles. Some evidence to support this assertion comes from the Kopriva et al. (2007) investigation, which randomized a limited set of accommodations students would receive while taking a traditional mathematics test. In that study, students who received accommodations the STELLA system recommended scored significantly higher than those who did not receive the STELLA recommended accommodations or only a subset. Further, students who received non-recommended accommodations or only a subset of the recommended ones scored the same as those who received no accommodations at all.

A possible explanation of the results in this study is that the expert ratings were based solely on information about the student gathered in the STELLA structured questionnaires as opposed to the broader knowledge of the student's teacher. The difference between what was focused on in STELLA and its subsequent recommendations and the teacher recommendations could be attributed to raters having a limited picture of the student. It is also possible that if the STELLA data collection forms had substantial flaws, both the ratings and the recommendations would be based on this flawed information, while perhaps being a poor fit to the students themselves. However, this explanation seems untenable because the analysis demonstrates that both sets of accommodations selected by the teacher were rated as no better than a random set of accommodations. If there would have been an important set of questions that was not asked on

the forms, it seems there would have been a systematic variation between teacher recommendations and the random set. However this did not occur.

Another competing explanation for the findings is that the rater training potentially biased the raters in favor of the accommodations recommended by STELLA. The rater training included instruction on the philosophy of matching large-scale test accommodations to the individual needs and strengths of the student as articulated in the introduction of this article. Thus, the raters were trained to consider test accommodations recommendations consistent with this philosophy. The STELLA decision taxonomy was also built on the same philosophy. However, at no point in the training were the raters taught any of the decision rules in the STELLA taxonomy. In fact, the study was designed and the raters trained by the first author, who understood the philosophy but never saw the STELLA decision taxonomy. Raters were blind to the source of the accommodation and gave the lowest rating to some recommendations made by STELLA, as evidenced in the descriptive statistics showing the minimum and maximum rating for each accommodation source. The experts on the panel gave their honest opinion that some accommodation sets recommended by STELLA were a poor manifestation of its philosophy. Replications of this study by independent researchers are always welcome and would provide further external validation. To date, however, the authors stand behind the raters and their findings based on the independence of the raters, the training procedures summarized above, and the non-significant rater by source interaction which found no evidence that individual raters were differentially biased towards any particular source of accommodation recommendation.

The teachers in this study had a tendency to recommend the same set of accommodations to meet the needs of all six of their students even though students differed widely in their individual profiles. In fact, teachers were specifically asked to choose diverse students to participate in the study, and the student profiles per teacher indicate that most teachers selected a reasonably heterogeneous group of students. In addition, the accommodations recommended by the teachers both before and after collecting relevant data were usually very similar, and subsequently received similar ratings for their appropriateness for a given student. Unfortunately, these characteristics of teacher behavior are consistent with other work that evaluates the robustness of test accommodations for students in special populations (see Kopriva & Lara, 2009; Kopriva & Koran, 2008). As an example, Plake and Impara (2006) found disability educators displayed great expertise in identifying the needs of different students but struggled to systematically match large-scale test accommodations to those needs. Douglas (2005) also reported that teachers could speak at length about their students' characteristics but could not link differences in their profiles to particular large-scale test accommodations. Along with the results discussed in this study, these findings suggest that the task of matching specific accommodations to individual needs, rather than the task of identifying individual needs, is where the teachers seem to be struggling when they are asked to recommend appropriate large-scale test accommodations. This supports a plausible and satisfying rationale for the results of previous studies suggesting that teachers predict at little above chance level which students with disabilities will benefit from particular test accommodations (Fuchs, et al., 2000; Helwig & Tindal, 2003, Weston, 2003). Further, this suggests that policies aimed at directing teachers to identify particular characteristics of their students and then link accommodations to them are not likely to be successful, as they do not disentangle the difficulties in matching appropriate large-

scale test accommodations from teachers' expertise in identifying the individual needs of their students. This is sobering when one considers that versions of this type of guidance are what is being used today by the two Race to the Top assessment consortia (Thurlow & Kopriva, 2015). However, a research-based accommodations decision algorithm could improve outcomes by complementing teachers' existing expertise in identifying the individual needs of their students.

The quantitative results here are remarkably robust. Teacher recommendations were often very similar across their six students, yet 95% of the explained variance in the ratings for all accommodation sets was attributable to students or source of the accommodation recommendation. This serves as solid evidence supporting the ability of the raters to distinguish differences in the appropriateness of the suggested accommodations for the individual students despite somewhat limited variability in the sets of accommodations across teachers. The non-significant rater by source interaction suggests that there is no differential rater bias toward or against any particular source of accommodation recommendation. Despite its limitations, this study provides noteworthy evidence to support further research that continues to pursue the design of computer-based systems for recommending appropriate test accommodations for ELLs.

Further research is warranted. A follow up study of the characteristics of students who had the poorest ratings for the accommodations recommended by the STELLA system may provide insight into areas for further research in recommending appropriate large scale test accommodations. Continued experimental studies with other accommodations, such as the one reported in Kopriva et al. (2007), while difficult to conduct, are also necessary for the continued verification of the decision trees behind the large-scale test accommodations recommendations.

References

Albus, D., Thurlow, M., Liu, K., & Bielinski, J. (2005). Reading test performance of english-language learners using an English dictionary. *Journal of Educational Research, 98*(4), 245-254.

Council of Chief State School Officers. (2003). *Annual survey of state student assessment programs: 2001-2002*. Washington, DC: Author.

Douglas, K. (2005). *Qualitative analysis of TTELL focus group data*. Unpublished manuscript, University of Maryland College Park.

Del Rosario Basterra, M., Trumbull, E. & Solano-Flores, G (2011, Eds.), *Cultural validity in assessment*. New York, NY: Routledge Publishers.

Fuchs, L., Fuchs, D., Eaton, S.B., & Hamlett, C.B. (2005). *Dynamic Assessment of Test Accommodations*. San Antonio, TX: PsychCorp.

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.

Helwig, R. and Tindal, G. (2003) An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children, 69*(2), 211–225.

Ketterlin-Geller, L.R. (2003). Establishing a validity argument for using universally designed assessments. Unpublished doctoral dissertation, University of Oregon, Eugene.

Kopriva, R.J., Carr, T., & Cho, M. (2006, June). *The selection taxonomy for English language learner accommodation.* Paper presented at the annual meeting of the Council of Chief State School Officers Large-Scale Assessment, San Francisco, CA.

Kopriva, R.J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues & Practice, 26*(3), 11-20.

Kopriva, R.J., & Koran, J. (2008). Proper assignment of accommodations to individual students. In R. J. Kopriva (Ed.), *Improving testing for English language learners* (pp. 217-254). New York: Routledge.

Kopriva, R.J., Koran, J., Hedgspeth, C. (2007) Addressing the importance of systematically matching students needs and test accommodations. In L. Cook and C. Cahahan (eds.), Large scale assessment and accommodation: What works? Arlington, VA: Council of Exceptional Children Press.

Kopriva, R.J., & Lara, J. (2009). Looking back and looking forward: Inclusion of all students in U.S.'s National Assessment of Educational Progress over the last 40 years and recommendations for the 21st century. In, *Celebrating the 50th Anniversary of NAEP*. USED Press, Washington, D.C.

Kopriva, R.J. & Myers, B. (2016). *A promising approach to developing student profiles and matching them to effective large-scale accommodations*. Manuscript under review.

Kopriva, R.J., Thurlow, M.L., Perie, M., Lazarus, S. S. & Clark, A. (in press). Test takers and the validity of score interpretations. *Educational Psychologist*.

Liu, K. K., Anderson, M.E., Swierzbin, 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.

Myers, B., & Kopriva, R.J. (2015). Decision trees linking individual student need to large-scale accommodations for English learners: A white paper. Retrieved from http://iiassessment.wceruw.org/research/

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practices,*

Plake, B., & Impara, J. (2006, March). Report from the Accommodation Station expert panel meeting. Savannah, GA. South Carolina Department of Education, Columbus, SC.

Rivera, C., & Collum, E. (2006). *State assessment policy and practice for English language learners: A national perspective*. Mahwah, N.J: Lawrence Erlbaum Associates.

Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Applied Measurement in Education, 27*(4), 236-247.

Thurlow, M.L. & Kopriva, R.J. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners, *Review of Research in Education, 39*(1), 331-369.

Tindal, G. (2006). The journey through the reliability of a decision-making model for testing students with disabilities. Presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.

Weston, T.J. (2003) *NAEP validity studies: The validity of oral accommodations in testing* (NCES 2003–06). Washington, DC: National Center for Education Statistics.

Table 1

Information Collected through Each of the Forms in the STELLA Data Collection Protocol

Records Form	Parent/Guardian Form	Teacher Form
Language of instruction	L1 ratings (3 point scale), 4 domains	English proficiency on a 4 point scale in 4 domains
English language proficiency test score (most recent)	Attendance in full-time academic programs in U.S.	L1 proficiency on a 4 point scale in 4 domains
L1 proficiency test score (most recent)	Length of time in U.S. schools	Perceived standardized score accuracy and judgments about reasons for inaccuracy
Type of ELL program	Consistency of attendance	Student's experience with standard test formats
	School atmosphere in native country if applicable	Student's understanding of the purpose of standardized testing
	Time (months, days/week, hours/day)	Classroom test condition options
	Number of students in classroom	Condition options that help student on classroom tests, evaluations
	School resources (e.g. chalkboards, desks, textbooks per student, other books, supplies for math or science, additional comments)	
	Types of assessments in native country	
	Grading practices	
	Test scores and experiences with testing in the native country	
	Test scores and experiences with testing in the US	

Table 2

Descriptive Statistics on Ratings by Source of Accommodation Recommendation

Source of Recommendation	Mean	Standard deviation	Minimum	Maximum	Mode	N
Teacher (before)	4.93	1.32	1	7	5	425
Teacher (after)	4.95	1.34	1	7	5	425
STELLA	3.67	1.49	1	7	4	425
Random	4.97	1.49	1	7	6	425