Educational Measurement
ISSUES AND PRACTICE
NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2020, Vol. 39, No. 4, pp. 65–75

Are Accommodations for English Learners on State Accountability Assessments Evidence-Based? A Multistudy Systematic Review and Meta-Analysis

Joseph A. Rios, Samuel D. Ihlenfeldt, and Carlos Chavez, *University of Minnesota*

The objectives of this two-part study were to: (a) investigate English learner (EL) accommodation practices on state accountability assessments of reading/English language arts and mathematics in grades 3–8, and (b) conduct a meta-analysis of EL accommodation effectiveness on improving test performance. Across all distinct testing programs, we found that at least one EL test accommodation was provided for both test content areas. The most popular accommodations provided were supplying students with word-to-word dual language dictionaries, reading aloud test directions and items in English, and allowing flexible time/scheduling. However, we found minimal evidence that testing programs provide practitioners with recommendations on how to assign relevant accommodations to EL test takers' English proficiency level. To evaluate whether accommodations used in practice are supported with evidence of their effectiveness, a meta-analysis was conducted. On average, across 26 studies and 95 effect sizes (N = 11,069), accommodations improved test performance by .16 standard deviations. Both test content and sampling design were found to moderate accommodation effectiveness; however, none of the accommodations investigated were found to have intervention effects that were statistically different from zero. Overall, these results suggest that currently employed EL test accommodations lack evidence of their effectiveness.

Keywords: effectiveness, English language learners, meta-analysis, systematic review, test accommodations

The United States has seen a large increase in the public school system of students who are unable to communicate and learn effectively in the English language (i.e., English learners [ELs]; U.S. Department of Education [USDOE], 2018; USDOE, n.d.). To ensure that these students are academically successful, the *Every Student Succeeds Act* (ESSA; 2015) stipulates that ELs are to be assessed in multiple content areas (mathematics, English language arts [ELA]/reading, and science), which will provide data to hold schools accountable for EL students' learning. However, the assessment of ELs presents fairness issues as these students may be put at a disadvantage as they cannot fully access the content and demonstrate their proficiency, leading to potential inaccuracies in score interpretations. To mitigate language serving as a source of construct-irrelevant variance (i.e., variance introduced by extraneous factors that are unrelated to the construct being assessed), accessibility to the test content can be increased by adapting the test format, administration, and/or response procedures in a manner that does not alter construct meaning (hereon referred to as test accommodations; American Educational Research Association [AERA] et al., 2014). In an effort to improve test fairness, ESSA (2015) stipulates that, "States must ensure that English learners are included in . . . statewide assessments by providing *appropriate* (italics added for emphasis) accommodations to all English learners" (U.S. Department of Education, 2016). An appropriate accommodation mitigates language as a source of construct-irrelevant variance (effective) by being sensitive to ELs' linguistic and educational background (relevant), while allowing for score comparability with test takers that did not receive an accommodation (valid; Abedi & Ewers, 2013; AERA et al., 2014).[1] Although multiple meta-analyses have demonstrated that EL accommodations in the literature are valid in that they do not provide an unfair advantage to EL test takers (see Kieffer, Lesaux, Rivera, & Francis, 2009; Li & Suen, 2012a), it is unclear if relevant accommodations are currently being provided to ELs and whether these accommodations are supported in the literature as being effective in increasing accessibility of the test content (hereon referred to as accommodation effectiveness). Bringing clarity to this issue is the overall objective of this paper.

Author contribution statement: The first author conceived of the presented idea. All authors were involved in data collection, authors 2 and 3 selected studies and coded variables, and the first author supervised all work. All authors conducted data analyses, interpreted findings, and drafted the article. The first and second authors conducted critical revisions of the article throughout the review process. Final approval of the version to be published was made by all authors.

Accommodation Relevance

There have been limited efforts to evaluate accommodation practices for ELs in state accountability testing (Abedi, Hofstetter, & Lord, 2004; Rivera & Collum, 2004; Rivera, Vincent, Hafner, & LaCelle-Peterson, 1997; Willner, Rivera, & Acosta, 2008). Across the few studies that have examined this issue, a number of trends have been observed. Specifically, prior to 2001, the majority of accommodations offered to ELs across states were both nonlinguistic (e.g., extra time and single-/small-group test administration) and also made available to students with disabilities, reflecting modifications that were not specifically designed for linguistic minorities (i.e., alone they do not increase language accessibility to the test content; Rivera et al., 1997; Rivera et al., 2000). In fact, Rivera and Collum (2004) found that across 46 states providing EL accommodations during the 2000–2001 academic year, 31 of the 75 (41%) available accommodations for ELs were designed within a taxonomy developed for students with disabilities (timing/scheduling, setting, presentation, and response), leading the authors to conclude that these accommodations were largely not suitable for ELs. This trend was found to continue after the passage of the No Child Left Behind Act (NCLB) as 64 of the 104 (62%) EL accommodations provided by all states in the 2006–2007 academic year were not designed for use with ELs (Willner et al., 2008).

Although these are disappointing results, there have been improvements in EL test accommodation availability over time in the United States. In 1994, only 52% of states provided EL test accommodations, while by 2007, this percentage increased to 100% (Rivera et al., 1997; Willner et al., 2008). Not only have the availability of accommodations improved exponentially, but we have also seen changes in the characteristics of accommodations, which reflect specific design for ELs. As an example, in reviews of accommodations during the 2000–2001 and 2006–2007 academic years, linguistic supports were found to be predominant across the distinct EL accommodations provided (61% in 2000–2001 and 98% in 2006–2007; Rivera & Collum, 2004; Willner et al., 2008). Of these linguistic accommodations, the majority of states were found to provide at least one non-English language test accommodation (e.g., test translation, side-by-side dual language test, directions in a non-English language). Though prior reviews provide an excellent overview of EL test accommodations offered nationwide, it is unclear: (a) whether the trends observed hold presently or improvements have been made as the last review was conducted over a decade ago (see Willner et al., 2008), and (b) if states provide guidance to practitioners on selecting accommodations that are relevant based on EL student characteristics (e.g., English language proficiency; ELP), as such relevance is necessary for an accommodation to be effective (Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007).

Accommodation Effectiveness

In the literature, accommodation effectiveness has been operationalized as performance improvement for ELs receiving the accommodation of interest when compared to ELs receiving no accommodation. To date, there have been three published meta-analyses that examined accommodation effectiveness for ELs.[2] Across these meta-analyses, there have been mixed findings. As an example, Kiefer et al. (2009) conducted a meta-analysis of 11 studies ($n = 38$ effect sizes; 17,455 native English speakers; 6,554 ELs), and found that the overall effect size across all test accommodations provided to ELs was .04*SD*.[3] Of the seven accommodations examined, only one—English language dictionaries and glossaries—had an overall positive effect on ELs' outcomes ($g = .18$; $n = 11$). Although Kiefer et al. (2009) found no significant moderators, Pennock-Roman and Rivera's (2011) meta-analysis (14 studies, 50 effect sizes) suggested that time limit may moderate accommodation effectiveness. Specifically, they found that when presented with restricted time limits, only pop-up English glossaries had a significant positive impact on test scores for ELs ($g = .29$; only based on two effect sizes), while the English dictionary/glossary test accommodation was most effective when little/no time constraints were present ($g = .23$; only based on three effect sizes). However, in the newest and largest meta-analysis conducted by Li and Suen (2012b; 19 studies, 85 effect sizes), accommodation effectiveness was moderated not only by time constraint, but also by English proficiency; EL students with low proficiency outscored their nonaccommodated counterparts by an average of .57 *SD*. Once accounting for this moderator, no significant differences between accommodation types were observed. Taken together, the results from these analyses do not provide a clear indication of test accommodation effectiveness for ELs as each meta-analysis had distinct inclusion criteria, analytic approaches, accommodation types studied, and moderators evaluated. Furthermore, both meta-analyses conducted by Kieffer et al. (2009) and Pennock-Roman and Rivera (2011) suffered from small sample sizes (i.e., fewer effect sizes) when making accommodation comparisons, which limits the validity of inferences made from these studies due to low statistical power.[4] Clearly, further empirical investigations are needed on this topic.

Study Objectives

As there has not been a review of accommodation practices since the passage of ESSA, the objective of this study is to survey the accommodations currently provided to ELs and evaluate whether the effectiveness of these accommodations is supported in the literature. This objective is investigated via two studies. In Study 1, we conduct a descriptive analysis of test accommodations for ELs taking state accountability measures in reading/ELA and mathematics for grades 3–8 during the 2017–2018 academic year. In doing so, the following research questions, which reflect those examined in research conducted prior to the passage of ESSA (e.g., Acosta, Rivera, & Willner, 2008; Rivera & Collum, 2004; Willner et al., 2008), are addressed:

1. What are the test accommodations provided to ELs on state testing programs of accountability? How do these accommodations differ across content areas (reading/ELA and math)?
2. How do currently available accommodations map to ELP levels? Do state testing programs provide recommendations for allowable accommodations based on test takers' ELP levels?

To investigate whether the effectiveness of the test accommodations found in Study 1 is supported by prior research, we conducted a meta-analysis of experimental research on EL test accommodations to address the following research question in Study 2:

3. Have the accommodations used in practice (found in Study 1) been empirically investigated in the literature?

If so, does research support the effectiveness of these accommodations? If not, are there alternative accommodations that have been studied that show effectiveness?

Results from these studies have the potential to inform testing programs about evidence-based accommodation practices that can be employed to improve the validity[5] of inferences concerning EL learning from accountability measures.

Study 1

Method

The sections that follow describe the data collection, variable coding (separated by each research question within Study 1), and interrater agreement processes of our systematic review of accommodations made available to ELs on state accountability assessments.

Data collection. Due to the creation of testing consortia Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC), certain states used the same assessment. As such, the unit of analysis for our systematic review is testing programs, which includes individual states and testing consortia.[6] Technical manuals were obtained for state accountability assessments of reading/ELA and mathematics in grades 3–8 across the 50 states in the continental United States (the District of Columbia and the U.S. territories were not included). In some cases, testing programs did not provide test accommodation information in their technical manuals, and instead possessed a separate test accommodation document. Regardless, all documents were obtained on state department of education websites, and in doing so, every effort was made to identify the most recent publicly available versions between September and November, 2018. In conducting our retrieval, a number of issues were identified. Specifically, Michigan did not provide a technical manual for their state assessment and Iowa provided an incomplete list of accommodations with little guidance on the usage of accommodations. As such, both Michigan and Iowa were excluded from the analysis, leaving a final sample of 29 testing programs (27 distinct state testing programs, SBAC, and PARCC).[7]

Accommodations provided. To examine test accommodations, we relied on the work of Acosta et al. (2008) and Kieffer et al. (2009) to provide a taxonomy of accommodations. Specifically, accommodations were grouped based on the interaction of presentation mode (written vs. oral/aural) and language (English vs. non-English language). Based on these references, a total of 32 accommodations were examined for their availability within each distinct testing program by content area (reading/ELA and math).[8] Furthermore, if an accommodation was provided in a non-English language, we examined which languages were made available. The reader is referred to Appendix A found online as Supporting Information to see the operational definitions for each accommodation.

Accommodation mapping to ELP levels. To evaluate how the accommodations currently provided to ELs match ELP levels, we relied on the recommendations set forth by Acosta et al. (2008). In their work, a group of experts in language testing, linguistics, second language acquisition, and educational measurement considered the level of ELP a student would need to benefit from a specific accommodation. This was done separately for accommodations provided in English and non-English languages. For the latter, experts also considered the native language proficiency of students when making recommendations by accounting for individuals' literacy and prior education. Across languages, accommodations were described as relevant for the following ELP levels: (a) beginning, (b) beginning and intermediate, (c) intermediate, (d) intermediate and advanced, (e) advanced, and (f) for any level. Although experts made a distinction with regard to whether an accommodation "may reduce" or is "likely to reduce" construct-irrelevant variance for ELs at a particular ELP level, this study focused only on the latter ("likely to reduce"). The reader is referred to Appendix B (online Supporting Information) to see how accommodations were mapped to ELP levels in this study. In addition, we examined whether testing programs provided guidelines to test administrators on allowable accommodations based on a test taker's ELP.

Interrater agreement. Interrater agreement was evaluated between the first and third authors for variable coding. This process consisted of first coding three testing programs together as part of coder training. The third author then coded all remaining testing programs, while the first author randomly coded approximately 40% (11 of 28). Agreement was then calculated for 52 variables separately using Cohen's kappa and percent agreement in the R package *irr* (Gamer, Lemon, Fellows, & Singh, 2012). Across variables, the median kappa value between the two raters was .74, while percent agreement was 86%. Any disagreements observed were resolved through consensus prior to conducting the final analysis.

Results

What are the test accommodations provided to ELs? The initial step to providing a snapshot of test accommodations across the country was investigating who was able to assign accommodations to ELs and how those accommodations were assigned. With regard to the former, for 46% of the 29 testing programs, accommodations were assigned by a team of educators who work with the student or are part of the student's IEP/504 plan team at the school level ($n = 12$), while 8% of testing programs made these designations at the school district level ($n = 2$; 12 testing programs did not provide this information). In investigating how accommodations were assigned, we found that 62% of testing programs specified that test accommodations must be aligned with those utilized for regular instructional assessments (nine were missing information for this variable). Below we provide a descriptive snapshot of testing program differences in accommodations by content area, which we detail in Appendix B (online Supporting Information).

Accommodations for math content area. All testing programs provided accommodations for math assessments with each offering an average of 8.69 ($SD = 3.39$, min $= 4$ [Tennessee, Ohio and Texas], max $= 16$ [Kentucky and Nebraska]) accommodations. The four most popular accommodations across testing programs were: (a) "flexible time/scheduling" (85%; $n = 22$); (b) "reading aloud test directions in English" (77%; $n = 20$); (c) "providing a commercial word-to-word dual language dictionary" (77%; $n = 20$); (d) "reading aloud test items in English" (73%; $n = 19$). In contrast, the following six accommodations were provided by only one testing program: (a) "simplified English test content;" (b) "allowing students to respond orally in English and transcribe their response;" (c) "use of a tape recorder to record test responses;" (d) "playing audio tape/CD of test directions in non-English language;" (e) "dual language test booklets;" (f) "dual language test questions for English passages." In terms of mode of presentation, all testing programs except Tennessee provided at least one oral/aural accommodation, and three testing programs—Ohio, Utah, and Wyoming—offered no written accommodations. Overall, nearly all programs provided an English (except for Florida and Tennessee) or non-English (except for Oklahoma) language accommodation with the latter being specified by 38% of distinct testing programs (for a list of non-English math accommodations by state, refer to Appendix C in the Supporting Information).[9]

Accommodations for reading/ELA content area. Across all testing programs, we found that accommodations for reading/ELA assessments were reflective of the ones observed for math assessments, although the average number of accommodations provided was slightly lower ($M = 8.40, SD = 3.27$, min $= 4$ [Ohio], max $= 16$ [Nebraska]). In addition, unlike mathematics assessments, not all accommodations were offered for reading/ELA, namely, "simplified English test content" ($n = 0$). The most popular accommodations for ELA, which closely resembled those provided in mathematics, included: (a) "providing a commercial word-to-word dual language dictionary" (85%; $n = 22$); (b) "flexible time/scheduling" (85%; $n = 22$); (c) "reading aloud test directions in English" (77%; $n = 20$); and (d) "reading aloud test items in English" (73%; $n = 19$). Similar to math, 16% of the 32 accommodations supplied for reading/ELA were included in only one testing program: (a) "allowing students to provide a written response in their native language;" (b) "playing audio tape/CD of test items;" (c) "playing audio tape/CD of test directions in a non-English language;" (d) "providing dual language test booklets;" (e) "providing dual language test questions for English passages." All but two testing programs, Ohio and Oklahoma, provided accommodations in a non-English language for test takers, while Florida and Tennessee did not offer English accommodations for ELA assessments. Across testing programs, the only non-Spanish accommodation provided for this content area was "read aloud test directions," which was offered in the following languages: Arabic, Chinese Mandarin, Navajo, Vietnamese, Polish, Portuguese, Haitian Creole, Russian, and Urdu. These languages for this accommodation were the same as those for the math content area. However, no oral/aural based accommodations were offered to ELs in Tennessee, while Ohio, Oklahoma, Utah, and Wyoming provided no written accommodations.

How do currently available accommodations map to EL proficiency levels?. It was atypical for testing programs to provide recommendations for assigning accommodations to suitable ELP levels ($n = 5$; neither consortium provided this guidance in their technical documentation). However, every testing program had at least one allowable accommodation for ELs at the beginner, intermediate, and advanced levels. The combinations of mode and language were found to differ by ELP level. Specifically, as expected, oral/aural non-English accommodations were recommended only for ELs with beginning ($n = 3$ accommodations) and beginning-intermediate ELP ($n = 18$) were made available, while the written English accommodations that were provided were solely recommended for EL test takers with intermediate-advanced ELP level ($n = 7$). In contrast, oral/aural English accommodations were offered across a wider range of ELP levels, including beginning ($n = 21$), beginning-intermediate ($n = 23$), intermediate ($n = 19$), and intermediate-advanced ($n = 7$) levels. Similarly, written non-English accommodations were made available for a broad set of EL test takers. Specifically, across the testing programs examined, these accommodations were offered to EL students with beginning ($n = 15$), beginning-intermediate ($n = 3$), intermediate-advanced ($n = 22$), and all ($n = 14$) ELP levels. These trends were found to be comparable across test content areas.

Next, we summarize the most popular accommodations observed across testing programs for each ELP level. In total, there were 16 accommodations coded that were allowable for beginning level ELP, with five of those being solely for beginners, and 11 for beginner-intermediate. The most utilized accommodation in our sample for the beginning ELP level was "read aloud test directions in English" (beginning-only, $n = 20$). Across testing programs, 19 intermediate accommodations were offered to ELs—11 of those were for beginner-intermediate, and four were for intermediate and intermediate-advanced each. Of these 19 accommodations, "allowing students to respond orally in English and transcribe their response" was the most used intermediate-only accommodation ($n = 13$). Although there were no accommodations allowable solely for the advanced ELP level, "providing a commercial word-to-word glossary" accommodation ($n = 22$) was made allowable at the highest rate for both the intermediate and advanced ELP levels. Finally, of the three accommodations allowable for all ELP levels, "extra time" ($n = 17$) was utilized at the greatest rate, followed by "electronic dual language pop-up glossary" ($n = 10$) and "customized dual language glossary" ($n = 10$).

Summary

Our systematic review showed that across testing programs, there was great variability in the overall number (e.g., across both content areas, Nebraska [16 accommodations] was found to offer four times more accommodations than Ohio [four accommodations]), language, and mode of accommodations provided. With regard to the latter, for mathematics and reading/ELA, there were a number of testing programs that did not provide non-English and/or oral-/aural-based accommodations, which may be of particular use to beginning-level ELP test takers. Furthermore, for those testing programs that did offer non-English accommodations, only New York, PARCC, and SBAC provided languages other than Spanish. Although 77% of ELs are native Spanish speakers (National Center for Education Statistics [NCES], 2018), approximately 1.1 million ELs speak a non-Spanish native language. Consequently, the only option for students outside of New York, PARCC, and SBAC is to utilize an English language accommodation, which may not be relevant for students who are incapable of understanding the language either orally, aurally, or in writing (i.e., beginning ELP test takers). This finding points to the lack of accessibility to fair testing practices for many EL students.

An additional finding is that although there may be many accommodations that are relevant for some ELs, only five testing programs were found to provide recommendations on allowable accommodations based on test-takers' ELP level. Therefore, for most testing programs, staff at the local school-level, who may lack formal training in educational measurement and/or may possess limited knowledge of the test fairness and accommodation literature, are placed with a great responsibility of choosing a number of accommodations (one or multiple) for students without best practice guidelines. Consequently, accommodations may not be assigned in a standardized fashion nor aligned with students' ELP level, which may undermine accommodation effectiveness (Kopriva et al., 2007). Due to the limited guidance provided to practitioners, the selection of relevant accommodations, and ultimately, the validity of inferences made from state accountability assessments about EL student learning, is put into question.

Study 2

The results from Study 1 indicate that practitioners are provided minimal direction in selecting relevant test accommodations for ELs on state accountability assessments. To provide evidence-based guidelines on accommodation practices, this study examined: (a) whether the accommodations currently employed in practice have been studied in the literature, and (b) of the studied accommodations, which are found to be most effective after controlling for subject and methodological factors using meta-analytic methodology. The current meta-analysis differs from previous studies on the topic in three ways (Kiefer et al., 2009; Li & Suen, 2012b; Pennock-Roman & Rivera, 2011). First, as this is an updated meta-analysis, it provides the most current and comprehensive sample of literature on the topic. For example, in contrast to the last study published on the topic (Li & Suen, 2012b), this meta-analysis increases the total sample of studies by 37% (26 total studies). Second, unlike previously published studies (Kiefer et al., 2009; Pennock-Roman & Rivera, 2011), this meta-analysis reduces the possibility of incorrect statistical inferences by controlling for effect size dependencies (see Scammacca, Roberts, & Stuebing, 2014) and considers sample size when making inferences concerning subgroup accommodation effectiveness. Third, this study is one of the first to account for research (e.g., sampling design and publication type) and accommodation (e.g., presentation mode) characteristics that have not previously been considered as potential moderators of accommodation effectiveness.

Method

Below we describe the search strategy, eligibility criteria, variable coding process, interrater agreement evaluation, and analyses conducted.

Search strategy. Primary studies on EL test accommodations were collected via three approaches (online Appendix F). The first strategy consisted of examining the references (i.e., backward citation searching) included in meta-analyses examining EL test accommodations published by Kieffer et al. (2009), Li and Suen (2012b), and Pennock-Roman and Rivera (2011). Second, the following databases were searched: (a) MNCat Discovery (aggregates from ERIC, PsycINFO, SAGE Premier, ProQuest, and Academic Search Premier) and (b) Google Scholar; (c) private organization (National Center on Educational Outcomes, WestED, CRESST) and testing company (ETS, SBAC, PARCC, WIDA, and College Board) research repositories. The included search terms were: ("English learner" OR "ELs" OR "limited English proficiency") AND ("linguistic modification" OR "test accommodation" OR "accommodation" OR "translation" OR "dictionary" OR "glossary" OR "simplified English" OR "accessibility features") AND ("test" OR "state test" OR "assessment"). Search results were limited to studies published as a journal article, technical report, and dissertation/thesis in the English language between 2010 and 2018. These years were chosen to account for a time period not investigated in the latest meta-analysis on EL test accommodations published by Li and Suen (2012b). This search was completed on September 18, 2018. Finally, both backward (using *Social Sciences Citation Index*) and forward (using Google Scholar) citation searching were conducted for studies found to meet the eligibility criteria (specified below) from our database search. This search process was completed on September 30, 2018.

Eligibility criteria. To be included, studies had to: (a) quantitatively examine the impact of one or multiple test accommodations on test performance for ELs, ELs with disabilities, or bilingual students in a K-12 educational context in the United States; (b) employ either a randomized control treatment, quasi-experimental, or single-group design; and (c) compare ELs receiving an accommodation to ELs receiving no accommodation if employing a between-subjects design. Studies were excluded if they: (a) did not evaluate test performance (e.g., examined accommodation usage; Roohr & Sireci, 2017); (b) examined accommodation effectiveness for a mixed sample of ELs and non-EL special education students in which the effect sizes for each subgroup could not be disaggregated (Cawthon, Leppo, Carr, & Kopriva, 2013); (c) did not provide information to calculate a standardized mean difference effect size (Cohen, Tracy, & Cohen, 2017); and (d) used a control group composed of non-ELs (see Abedi et al., 1998).

Variable coding. Beyond sample and effect sizes, a total of eight moderator variables were coded for, which were categorized as subject, research, and accommodation characteristics. The identification of these variables was influenced by prior meta-analyses of test accommodations as well as methodological factors that have been found to impact meta-analyses in education (Cheung & Slavin, 2016; Li & Suen, 2012b; Pennock-Roman & Rivera, 2011). Below we present a description and rationale for the inclusion of these variables. The coding protocol for this study, which includes an operationalization and coding strategy for each variable, can be found in Appendix D in the online Supporting Information.

Subject characteristics. Two variables related to subject characteristics, grade and ELP level, were examined as moderators (similar to Li & Suen, 2012b). Grade was coded dichotomously as K-6 versus grade 7 and above (reference group). As most EL students in younger grades are assumed to on average possess lower ELP, it was hypothesized that accommodation effectiveness would be significantly greater for K-6 subjects. ELP level was dichotomously coded as low ELP versus moderate/high or mixed (i.e., a subject pool consisting of multiple ELP levels; reference group) ELP based on primary authors' descriptions of the subject pool (similar to Li & Suen, 2012b). If ELP level was not described, it was assumed that participants were moderate/high or mixed ELP. We hypothesized that accommodation effectiveness would be significantly greater for studies with low ELP subjects (Li & Suen, 2012b).

Research characteristics. Test content, sampling design, and publication type were included as moderators of research characteristics. Test content was coded dichotomously as math/science versus other (reference group; similar to Li & Suen, 2012b). This was done as students in the United States have been found to struggle on STEM-related assessments, which led us to hypothesize that even if test accessibility is improved, low content knowledge may attenuate performance differences between treatment and control groups (NCES, 2018). Additionally, as noted by Charness, Gneezy, and Kuhn (2012), study design (within-subject versus between-subject) has been found to moderate findings in the psychological literature. To account for this, we dichotomously coded sampling design as within- versus between-subjects designs (reference group), and hypothesized that the former would have significantly larger differences. Finally, publication type (peer-reviewed journal articles versus gray literature) was included as a moderator as the literature published in peer-reviewed journals has been found to have substantially larger intervention effects than gray literature (reference group; Cheung & Slavin, 2016).

Accommodation characteristics. Presentation mode, language, and type of accommodation were included as moderator variables. Presentation mode was dichotomously categorized as oral/aural versus written (reference group), while accommodation language was dichotomized as non-English versus nonlinguistic/English (reference group). No hypothesis regarding accommodation effectiveness was made for presentation mode; however, it was hypothesized that non-English language accommodations would be more effective. This hypothesis was based on the assumption that, if the non-English accommodation was provided in the native language of students, it would increase accessibility to the test content. It is recognized that such an assumption may not hold for individuals who have not received formal reading and/or writing instruction in their native language and are receiving an accommodation that requires them to read and/or write. However, over 40% of ELs in grades 6–12 are foreign-born (Sugarman & Geary, 2018), which may suggest that many EL students from middle to high school have received some form of educational instruction in a non-English language.[10] Finally, four accommodations were included in this study: (a) test translation (combined dual language test book and test translation/adaptation; reference group); (b) simplified English; (c) use of dictionaries/glossaries (combined English dictionary/glossary, dual language dictionary, and picture dictionary); (d) combined accommodations (i.e., employing two or more accommodations simultaneously).[11] These accommodation types were examined as they each possessed at minimum 10 effect sizes, which is the minimum required for the inclusion of a moderator variable (Higgins & Green, 2011). As this criterion was not met by the extra time ($n = 3$), read aloud ($n = 2$), and pictorial aide ($n = 2$) test accommodations, they were dropped from the moderator analysis. No hypothesis was made with regard to accommodation-type differences.

Interrater agreement. Interrater agreement was calculated for every variable separately using Cohen's kappa and percent agreement in the *R* package *irr* (Gamer et al., 2012) based on the first author randomly double coding approximately 20% of articles (5 of 26 studies; all articles were coded by the second author). Across all 13 variables (including sample sizes, means, standard deviations, and moderators) and 12 independent samples (some studies produced more than one sample), the average κ was .71 and the average percent agreement was 85% (ranged from 75% to 92%). Rater disagreements were settled by consultation between the first two authors.

Analyses. Upon coding the means and standard deviations of the treatment conditions for each study, standardized mean difference effect sizes were computed based on Cohen's d formula for a between-subjects design:

$$d = \frac{y_1 - y_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}, \tag{1}$$

where y_1 and y_2 are sample means, n_1 and n_2 are sample sizes, and S_1^2 and S_2^2 are the standard deviations for the experimental and control groups, respectively. For studies that employed a single-subject design (13 effect sizes from six studies), Cohen's d was computed as:

$$d = \frac{(X_2 - X_1)\sqrt{2(1-r)}}{\sqrt{S_1^2 + S_2^2 - 2rS_1S_2}}, \tag{2}$$

where X_2 is the experimental mean, X_1 is the control mean, S_2^2 is the experimental standard deviation, S_1^2 is the control standard deviation, and r is the correlation between the control and experimental scores. When not provided by the primary study, the correlation between pretest and posttest was imputed to be equal to .70 based on a sensitivity analysis showing no significant difference in results for values that ranged from .50 to .90. Due to the tendency for Cohen's d to slightly overestimate effect sizes when small samples are present, all effect sizes were converted to Hedges' g to account for any potential biasing based on the following conversion:

$$g = j(df)d, \tag{3}$$

where $j(df)$ is a correction factor equal to $1 - \frac{3}{4(n_1+n_2-2)-1}$. These calculations were completed in the *R* package *compute.es* (Del Re, 2013).

Prior to calculating the mean effect size[12] and effect size heterogeneity across all studies, diagnostic analyses of outliers and publication bias were conducted. With regard to outliers, any effect size found to be greater than three standard deviations from the median effect size was downweighted to be equal to three standard deviations. Publication bias was examined via the funnel plot procedure and Duval and Tweedie's (2000) trim-and-fill method. To account for effect size dependencies due to a single primary study producing multiple effect sizes, the robust variance estimation (RVE) procedure with hierarchical effects (i.e., effect sizes were nested within studies) was implemented in the *R* package *robumeta* to obtain accurate variance estimates based on a method-of-moments estimator (Fisher, Tipton, & Hou, 2016; Hedges, Tipton, & Johnson, 2010). The average effect size and heterogeneity of effect size estimates were calculated using a random-effects intercept-only model based on restricted maximum likelihood estimation. In both this model and the moderator model (described below), an inverse variance weight was applied to the effect sizes. Heterogeneity was assessed via the I^2 statistic proposed by Higgins and Thompson (2002). This statistic reflects the proportion of the variation in effect size estimates due to heterogeneity as opposed to chance and provides a practical interpretation of heterogeneity in which $I^2 < 50\%$ is representative of small heterogeneity, $50\% \leq I^2 < 75\%$ is representative of medium heterogeneity, and large heterogeneity is represented by an $I^2 \geq 75\%$ (Higgins & Thompson, 2002).

Next, accounting for the moderators noted above, a comparative analysis of accommodation type was conducted via the following random-effects meta-regression model using restricted maximum likelihood estimation:

$$\begin{aligned}\hat{y} = {} & b_0 + b_1\,(grade) + b_2\,(ELP\ level) \\ & + b_3\,(test\ content) + b_4\,(sampling\ design) \\ & + b_5\,(publication\,type) + b_6\,(presentation\,mode) \\ & + b_7\,(language\ of\ accommodation) \\ & + b_8\,(test\,translation) + b_9\,(simplified\ English) \\ & + b_{10}\,(use\ of\ dictionaries\,/glossaries) \\ & + b_{11}\,(combined\,accommodations) + e,\end{aligned} \tag{4}$$

where $\hat{y}$ was equal to Hedge's g of accommodation effectiveness, b_0 was equal to the average effect size for the dependent variable of interest after controlling for all included variables (not of substantive interest), and e was the residual term. All other variables were entered based on their coding schemes described in the "variable coding" section. After controlling for moderator variables, an F-statistic was calculated to test the equality of estimates for the intervention types using a sandwich estimator for the variance–covariance matrix and a small sample correction for the p value via the *clubSandwich R* package (Pustejovsky, 2019). If significant at $p < .05$, post-hoc multiple-contrast hypotheses were conducted. To control for Type I error, the Bonferroni procedure was employed (Dunnett, 1955).

Results

After removing duplicates, our database and citation searching produced 2,219 references, which underwent title and abstract screening. Of these, full-text screening was conducted for 76 studies, and 26 (k) met our eligibility criteria for inclusion. This final sample produced 95 effect sizes (N) based on 11,069 EL test takers (see Appendix E in online Supporting Information for a list of these references). The sampled studies were written by 19 distinct first authors and published between 1998 and 2018. Only 27 of the 95 effect sizes (28%) came from studies published in peer-reviewed journals.

Average effect size and heterogeneity. Prior to computing the average effect size, both outliers and publication bias were evaluated. In investigating the former, two outliers were identified, and were thus downweighted to be equal to three standard deviations from the median effect size. In terms of publication bias, both the funnel plot (see Appendix G in online Supporting Information) and trim-and-fill (the estimated number of missing studies on the left side of the distribution was approximately 0 [$SE = 5.59$]) methods demonstrated that effect size estimates were scattered symmetrically across the median effect size across all studies, suggesting no presence of publication bias. Once accounting for these diagnostic analyses, the average effect size and effect size heterogeneity were calculated. Across all studies, test scores improved by an average of .16 *SD* ($SE = .06$; 95% CI: .04, .28) when ELs were provided test accommodations; though a large degree of heterogeneity was noted within the sample ($I^2 = 90.72\%$), indicating the need for a moderator analysis.

Moderator analysis. Based on 24 unique studies and 88 effect sizes (seven effect sizes were dropped from the moderator analysis due to accommodations with small sample sizes), the inclusion of the outlined moderators (minus accommodation presentation mode due to a lack of variability in primary studies) accounted for an additional 11% of variance when compared to the null model (i.e., not including moderators). Model results are presented in Table 1.

Subject characteristics. With regard to grade level, 11 of 24 studies included at least one K-6 sample, which accounted for 36% of effect sizes. However, no significant difference in accommodation effectiveness was observed when comparing samples in K-6 and grade 7 and above ($\beta = .13, p = .10$). In terms of sample ELP level, 22% of effect sizes were attributable to low ELP samples (4 of 24 studies included at least one low ELP sample). Similar to grade level, no significant difference was found between ELP level ($\beta = .23, p = .15$).

Research characteristics. The majority of studies investigated accommodation effectiveness for assessments that possessed either science or math content (21 out of 24 studies; 75% of effect sizes). Model results demonstrated that accommodation effectiveness was lower by .40 *SD* ($p < .05$) for science/math assessments when compared to those assessing other content (reading and history). Similarly, sampling design was found to be a significant moderator. That is, when comparing within-subjects ($k = 7$, $n = 21$) and between-subjects ($k = 17$, $n = 67$) designs, the former was found to be significantly larger by .36 *SD* ($p < .05$). However, no significant differences were observed between published and gray literature ($\beta = -.18, p = .12$).

Accommodation characteristics. Across studies, a great imbalance in the presentation mode of accommodations was observed. Specifically, 93% of all language-based accommodation effect sizes were investigated for the written format. Due to this lack of variability in primary studies, presentation mode was dropped from the moderator analysis. With regard to language of presentation, English was found to be the predominant language for the accommodations in the sampled studies (51%), followed by non-English (38%), nonlinguistic (7%), and a mix of language-type (5%) accommodations. In comparing non-English versus English/nonlinguistic

Table 1. Moderator Analysis for EL Test Accommodations

	Accommodation-Only Model (k = 24, n = 88)I^2 = 91.60; τ^2 = .15			Moderator Model (k = 24, n = 88)I^2 = 79.95; τ^2 = .07		
Moderator	**Estimate**	**S.E.[a]**	**95% CI**	**Estimate**	**S.E.[a]**	**95% CI**
Intercept[b]	.09	.09	−.11, .30	.06	.19	−.39, .52
Subject Characteristics						
Grade[c]	–	–	–	.13	.09	−.07, .34
ELP Level[d]	–	–	–	.23	.19	−30, .77
Research Characteristics						
Sampling Design[e]	–	–	–	.36*	.14	.05, .67
Test Content[f]	–	–	–	−.40*	.15	−.82, .01
Publication Type[g]	–	–	–	.18	.15	−.15, .52
Accommodation Characteristics						
Presentation Mode[h]	–	–	–	–	–	–
Language[i]	–	–	–	.10	.10	−.20, .39
Simplified English[j]	−.04	.11	−.27, .19	.11	.15	−.30, .53
Dictionary/Glossary[j]	.06	.19	−.34, .45	.14	.15	−.21, .49
Combined Accommodations[j]	.34	.14	−.16, .84	.09	.23	−.54, .72

Note. Caution should be given when interpreting the parameter estimates for the combined accommodation type due to low degrees of freedom. τ^2 is equal to the between-study variance (i.e., the degree of variance in effects observed in different studies). *p < .05; **p < .01.

[a] Robust standard errors are provided based on the robust variance estimation procedure.

[b] In the accommodation-only model, the intercept is equal to the average improvement in the performance for the test translation accommodation. The intercept in the moderator analysis is interpreted as the average effect size for independent samples that are published in peer-reviewed journals, included a sample in grade 7 or above with medium/high or mixed ELP level, employed a between-subjects design, examined accommodation effectiveness on a non math/science assessment, and included a test translation accommodation.

[c] *Grade* was dichotomously coded (reference was grade 7 and above).

[d] *ELP level* was dichotomously coded (reference was medium/high or mixed ELP level).

[e] *Sampling design* was dichotomously coded (reference was between-subjects design).

[f] *Test content* was coded dichotomously (reference was non math/science).

[g] *Gray literature* was dichotomously coded (reference was peer-reviewed journal articles).

[h] *Presentation mode* was dropped from the analysis due to the lack of variability found.

[i] *Language* was dichotomously coded (reference was English/nonlinguistic).

[j] These accommodation types were dummy coded with the test translation accommodation serving as the reference group.

Based on a prior hypothesis, a one-tailed statistical test was employed for the following moderators: (a) grade, (b) ELP level, (c) sampling design, (d) test content, (e) publication type, and (f) language.

accommodations, no significant difference was observed ($\beta = .10, p = .20$).

A total of 14 distinct test accommodations were investigated (once dropping extra time, real aloud, and pictorial aide accommodations); however, three major accommodation types accounted for 86% of all effect sizes. These accommodation types were: (a) simplified English (n = 27); (b) test translation (combined test translation/adaptation [n = 16] and dual language test booklets [n = 8]; n = 24); (c) use of dictionaries or glossaries (combined English [n = 13], dual language [n = 6], and picture [n = 6]; n = 25). The remaining 14% of effect sizes were attributed to combinations of the accommodations listed above.[13] In using test translation as the reference accommodation, no significant differences were found for simplified English ($\beta = .11, p = .50$), providing dictionaries/glossaries ($\beta = .14$, $p = .36$), or combining accommodations ($\beta = .09, p = .71$) when accounting for moderators. Furthermore, the unadjusted mean effect size (i.e., not controlling for moderators) for each accommodation type was found to be not statistically different from zero, indicating that the accommodations examined were, on average, ineffective in improving test performance (see Table 1).

Discussion

This meta-analysis demonstrated that the literature has predominantly investigated accommodations provided in the written format (less than 10% of effect sizes examined an oral-/aural-based accommodation) with three accommodations accounting for 86% of available effect sizes: (a) simplified English, (b) test translation, and (c) use of dictionaries or glossaries. However, as found in Study 1, in practice, limited testing programs employ simplified English (mathematics: Kentucky; reading/ELA: no testing program) or test translation, which in this study included test translation/adaptation (mathematics: PARCC, New York, and Pennsylvania; reading/ELA: New York and Pennsylvania) and dual language test booklets (mathematics: Pennsylvania; reading/ELA: Pennsylvania). This suggests that there is a major disconnect between the accommodations that have been studied for their effectiveness and those that are currently being provided on state accountability assessments. Thus, as noted by Abedi and Gándara (2006), it appears that accommodation practices are not based on empirical research.

The one accommodation employed in practice (see online Appendix B) and investigated in the literature across numerous studies (n = 27) is the use of dictionaries/glossaries. However, similar to Li and Suen (2012), our meta-analysis found no support for any accommodation (including use of dictionaries/glossaries) having an effect significantly different from zero. One potential reason for the nonsignificant findings is that each accommodation type examined possessed a large degree of uncertainty due to primary authors investigating accommodation effectiveness with very small sample sizes (51% of effect sizes included in the moderator analysis were based on a total sample size of less than 100). Regardless, it appears that there is little evidence to indicate that the current accommodation approaches improve test content accessibility for linguistic minorities.

Summary and Concluding Discussion

This paper demonstrates that the majority of current approaches to accommodating linguistic minorities on state accountability measures have not been empirically investigated in the literature for their effectiveness. Although it is possible that these accommodations may be effective, it has yet to be demonstrated. Of those accommodations that have been studied, none have been found to be associated with performance improvements that are significantly different from zero. Thus, there may be serious threats to test fairness on state accountability measures for ELs as there appears to be no evidence for effective strategies to mitigate language serving as a source of construct-irrelevant variance. As a consequence, students may be unable to fully demonstrate their standing on the construct being measured due to an inability to access the test content because of language, which may lead to invalid inferences concerning EL student learning. This, in turn, may potentially undermine U.S. accountability efforts to improve EL educational outcomes.

Limitations

In general, the findings from this paper are limited to accommodation practices within the continental United States and cannot be generalized to assessing linguistic minorities in other countries or multinational contexts. As both studies relied on systematic review processes, they are limited by the search strategies employed. Specifically, in Study 1, our findings are limited to the information provided in the technical manuals investigated, and thus, may not accurately reflect test program practices that were not included in the technical documentation reviewed. In Study 2, a concerted effort was made to conduct a thorough literature review by employing multiple search strategies to obtain all available research including gray literature. Yet, it is possible that some studies were missed due to not including certain search strategies (e.g., professional research organization listservs) or because they may not have been made publicly available (e.g., studies conducted by school districts). Additionally, in Study 2, due to sample size restrictions, accommodation types had to be collapsed into categories that were theoretically alike, which may have led to ignoring important heterogeneity. As an example, dictionaries used as an accommodation may vary in complexity, quality, and alignment to students' ELP. Similarly, in practice, students may receive any number of different possible combinations of accommodations; however, we were unable to examine the effectiveness of these combinations as there was not enough primary research to allow for a sufficiently powered analysis. Beyond sample size concerns, primary authors often did not provide sufficient detail to disaggregate heterogeneity within an accommodation type. Thus, we may have been unable to control for accommodation characteristics that may have influenced effectiveness. Similarly, limited information was specified on how EL classification was defined. This lack of specificity is not trivial, as research on accommodation effectiveness may be affected by the contexts in which accommodations are given, which include the population of test takers. This potential source of error could not be examined in our meta-analysis; however, EL designation's role as a source of variability should not be ignored in future research. Another limitation of our study is that we only investigated accommodation effectiveness and relevance, two of four factors in the discussion of the appropriateness of accommodations (Abedi & Ewers, 2013). Accommodation effectiveness ought not to be confused with accommodation validity as it is possible to mitigate language as a source of construct-irrelevant variance at the cost of changing the underlying construct. Thus, it is possible for an accommodation to be effective but still not be valid because it provides an unfair advantage to ELs. Although prior research has demonstrated that accommodations allow for fair comparability between students who are or are not accommodated (Kieffer et al., 2009; Li & Suen, 2012a), ultimately, more research must be conducted to compare the performance of non-ELs who are accommodated and non-ELs who are not accommodated.

Research implications

In spite of these limitations, the findings from this study point to the need for the measurement community to assist in building a solid experimental research based on accommodation effectiveness to provide practitioners with best practice guidelines. This is vital as our review demonstrated that minimal work has been conducted in this area since NCLB was passed (only 20 experimental studies have been conducted since 2001). To entice engagement in this research area, there is a need for funding agencies, such as the Institute of Education Sciences, state departments of education, and private organizations, to submit specific calls for research on the development and evaluation of EL test accommodations as this type of research requires both extensive time and financial resources.

In investigating accommodations, the field should shift from asking, "Is a particular accommodation effective?" to "For whom, and under what conditions, is a particular accommodation effective?" One way to approach this from a subpopulation standpoint is to recognize EL heterogeneity and begin to study specific EL subpopulations (e.g., a low ELP subpopulation that are illiterate in their native language) by collecting large sample sizes that account for idiosyncrasies and allow for sufficiently powered analyses. Our meta-analysis showed that this would not be feasible in prior research as some studies had as few as six participants (median $= 96, SD = 158.73$). Furthermore, in evaluating the effectiveness of accommodations, funded research should consider the impact of research design as our meta-analysis demonstrated a significant difference between within- and between-subject designs. Although each design possesses its own advantages and disadvantages, the within-subjects design with counterbalancing has been suggested to be more powerful in identifying true differences in treatment effects (Rosenthal & Rosnow, 2008). Second, in considering potential contextual effects in accommodation effectiveness, there is a need to consider how accommodations may interact with the diversity of language demands, item types, and response formats that are present on state accountability measures. Third, we suggest that the current operational definition of accommodation effectiveness (performance improvement when comparing accommodation and control groups) be revisited as it is possible that an accommodation is effective without leading to performance differences.

Therefore, it is recommended that studies should also incorporate evidence around test content accessibility, which can be gathered via posttest surveys, qualitative interview data, and think-aloud protocols. Regardless of the approach, we must move beyond simple test score comparisons to fully understand test accommodation effectiveness.

Implications for practice

Due to the limited evidence of effective accommodation practices, it is recommended that educational stakeholders should use caution when interpreting EL's scores from state accountability measures as language may serve as a serious validity threat. Additionally, as many states require an alignment between classroom and state accountability assessments, classroom practitioners should consider the need to assign accommodations based on the alignment across accommodation characteristics, test content, as well as students' prior education, and native and English language proficiencies. Beyond that, school staff can assist in improving the current state of evidence around accommodation practices by conducting local research projects with their students. For example, as many testing programs employ the same accommodations across all assessments, school staff may have the opportunity to: (a) perform qualitative interviews with students after taking classroom exams to ascertain if the administered accommodations are improving accessibility of the English language content; (b) design experimental studies with EL students by randomly assigning some students to an accommodation condition and others to a control condition when taking a low-stakes classroom exam (i.e., there are no personal consequences for the test taker associated with their performance). If such research is conducted, it is recommended that school staff share their findings with other practitioners and researchers through publicly available repositories (e.g., EdArXiv). These efforts will assist the field in building a large collection of primary research, which can serve as the foundation for creating evidence-based reform of EL assessment practices. By doing so, we may be able to assist policymakers in making valid score-based inferences concerning EL student performance on state accountability measures.

Notes

[1] Abedi and Ewers (2013) also stipulated that accommodations should be logistically implementable (i.e., feasible). However, as this is dependent on the resources of every testing program, it is not a focus of this paper.

[2] Chiu and Pearson (1999) were technically one of the first to review the effectiveness of EL test accommodations; however, their sample size consisted of only five effect sizes limiting the generalizability of their study.

[3] This meta-analysis was based on initial results published in a report by Francis, Rivera, Lesaux, Kieffer, and Rivera (2006).

[4] Although the overall number of effect sizes was larger in Li and Suen (2012), it is unclear how many effect sizes were analyzed for each accommodation in their paper.

[5] This refers to the validity of inferences made from assessments given to ELs, which is not the same as the validity of the accommodation.

[6] At the time of coding, there were two test consortia, SBAC and PARCC. SBAC consisted of 13 states (CA, CT, DE, HI, ID, MT, NV, ND, OR, SD, VT, WA, WV), while PARCC consisted of eight (CO, IL, LA, MD, MA, NJ, NM, RI).

[7] The years of collected technical manuals are as follows: VA (2015); WY (2016); AL, AK, AZ, AR, SBAC, PARCC, FL, IN, KS, KY, ME, MN, MS, NE, NH, NY, OH, PA, SC, TN, TX, WI (2017); NC, UT (2018); GA (2019).

[8] Flexible time/scheduling was included in our analyses—Rivera and Collum (2004) treated this accommodation as an EL sensitive accommodation so we have included it, although this in contrast to Acosta, Rivera, and Willner (2008).

[9] As noted by our reviewers, there are other states that may provide native language accommodations not observed in our review of technical manuals. For instance, it was pointed out that Minnesota offers Somali, Hmong, and Vietnamese language accommodations, and Texas offers testing in Spanish. Additionally, a reviewer pointed out Michigan provides accommodations in Arabic. We did not find this information provided in the technical manuals sampled; however, we acknowledge that testing programs could have offered accommodations not described in the technical manuals.

[10] This estimate of students who are foreign-born comes from research by the Migration Policy Institute based on U.S. Census Bureau pooled data from the 2012–2016 American Community Survey. The national estimate for students in K-5 is much lower (approximately 18%). However, as noted by one reviewer, estimates of EL students who are foreign-born may vary depending on the source, grade band, state, and EL criteria.

[11] Li and Suen (2012) excluded some studies that employed multiple accommodations.

[12] A prospective power analysis was conducted to estimate the number of effect sizes needed to yield an average effect size of .10 and .25 (what works clearinghouse criteria for a substantial effect) with 80% power assuming an average sample size of 150 participants, moderate heterogeneity, and a two-tailed alpha level of .05. Based on these assumptions, this analysis demonstrated that for average effect sizes of .10 and .25, a total of 42 and 7 effect sizes is needed to attain 80% power, respectively.

[13] These combinations were as follows: bilingual glossary and oral presentation of test content ($n = 1$); bilingual glossary and picture dictionary ($n = 1$); English dictionaries/glossaries and extra time ($n = 1$); picture dictionary and oral presentation of test content in English ($n = 1$); oral presentation of test content in English, picture dictionary, and bilingual glossary ($n = 1$); test translation and oral presentation of test content in non-English language ($n = 1$); simplified English and test translation ($n = 2$); and simplified English and picture dictionary ($n = 4$).

Acknowledgment

The authors would like to thank Dr. Kristin Liu, Senior Research Associate at the National Center on Educational Outcomes, for providing assistance with the literature search conducted in Study 2.

References

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced Assessment Consortium. http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/08/Accommodations-for-under-represented-students.pdf

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*(4), 36–46.

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*, 1–28.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Technical Report 478). National Center for Research on Evaluation, Standards, and Student Testing.

Acosta, B. D., Rivera, C., & Willner, L. S. (2008). *Best practices in state assessment policies for accommodating English language learners: A Delphi study*. Arlington, VA: Center for Equity and Excellence in Education, the George Washington University.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cawthon, S., Leppo, R., Carr, T., & Kopriva, R. (2013). Toward accessible assessments: The promises and limitations of test item adaptations for students with disabilities and English Language Learners. *Educational Assessment*, *18*, 73–98.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*, 1–8.

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*, 283–292.

Chiu, C. W. T., & Pearson, D. P. (1999, June). Synthesizing the effects of test accommodations for special education and limited English proficient students [Paper presentation]. National Conference on Large Scale Assessment, Snowbird, UT, United States.

Cohen, D., Tracy, R., & Cohen, J. (2017). On the effectiveness of pop-up English language glossary accommodations for EL students in large-scale assessments. *Applied Measurement in Education*, *30*, 259–272.

Del Re, A. C. (2013). *compute.es: Compute effect sizes.* R package version 0.2-2. http://cran.r-project.org/web/packages/compute.es

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, *50*, 1096–1121.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455–463.

ESSA. (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114–95 § 114 Stat. 1177 (2015-2016). https://www.ed.gov/essa?src=rn

Fisher, Z., Tipton, E., & Hou, Z. (2016). *Robumeta: Robust variance meta-regression*. R package version 1.8. https://cran.r-project.org/web/packages/robumeta/robumeta.pdf

Francis, D., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *irr: Various coefficients of interrater reliability and agreement*. R package version 0.84. https://CRAN.R-project.org/package=irr

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65.

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0). London, UK: Cochrane Collaboration.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*, 1168–1201.

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 and Practice*, *26*, 11–20.

Li, H., & Suen, H. K. (2012b). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education*, *25*, 327–346.

Li, H., & Suen, H. K. (2012a). Are test accommodations for English language learners fair? *Language Assessment Quarterly*, *9*, 293–309.

National Center for Education Statistics. (2018). *The nation's report card: 2017 mathematics results*. https://www.nationsreportcard.gov/reading_math_2017_highlights/files/infographic_2018_math.pdf

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 Practice*, *30*, 10–28.

Pustejovsky, J. (2019). clubSandwich: Cluster-robust (sandwich) variance estimators with small sample corrections. R package version 0.3.5. https://CRAN.R-project.org/package=clubSandwich

Rivera, C., & Collum, R. (2004). *An analysis of state assessment policies addressing the accommodation of English language learners*. Washington, DC: Center for Equity and Excellence in Education, the George Washington University.

Rivera, C., Stansfield, C. W., Scialdone, L., & Sharkey, M. (2000). *An analysis of state policies for the inclusion and accommodation of English language learners in state assessment programs during 1998–1999*. Washington, DC: Center for Equity and Excellence in Education, the George Washington University.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). *Statewide assessment programs: Policies and practices for the inclusion of limited English proficient students*. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation, The Catholic University of America.

Roohr, K. C., & Sireci, S. G. (2017). Evaluating computer-based test accommodations for English Learners. *Educational Assessment*, *22*, 35–53.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). New York: McGraw-Hill.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, *84*, 328–364.

Sugarman, J., & Geary, C. (2018). *English learners in Minnesota: Demographics, outcomes, and state accountability policies*. Migration Policy Institute. https://www.migrationpolicy.org/research/english-learners-demographics-outcomes-state-accountability-policies

U.S. Department of Education. (n.d.). *Our nation's English learners: What are their characteristics*? https://www2.ed.gov/datastory/el-characteristics/index.html

U.S. Department of Education. (2016). *Fact sheet for final regulations: Title I, part A and part B*. https://www2.ed.gov/policy/elsec/leg/essa/essaassessmentfactsheet1207.pdf

U.S. Department of Education, National Center for Education Statistics. (2018). *English language learners in public schools*. https://nces.ed.gov/programs/coe/indicator_cgf.asp

Willner, L. S., Rivera, C., & Acosta, B. D. (2008). *Descriptive study of state assessment policies for accommodating English language learners*. Washington, DC: Center for Equity and Excellence in Education, the George Washington University.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website: https://onlinelibrary.wiley.com/doi/10.1111/emip.12337