Exploring Validity and Fairness of the Text-to-Speech Accommodation for College and Career Readiness Assessments

Elizabeth Stone, Cara Laitusis, and Carlos Cavalie

Educational Testing Service

The PARCC Theory of Action and Impetus for the Current Study

The Partnership for Assessment of Readiness for College and Careers (PARCC) is a group of states working together to develop and implement assessments in English language arts/literacy and mathematics. Through this program, the consortium sought to develop a robust assessment system that would provide timely information about student performance to relevant stakeholders that supports and promotes the development of effective K–12 instruction, with the goal of preparing students for postsecondary education and the workforce. The theory of action for the assessment system included several key design elements: incorporation of and alignment to the Common Core State Standards (CCSS), which are shared by PARCC states (CCSS; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010a, 2010b); increased use of technology to deliver innovative item types and tasks to enrich measurement of student proficiency by capturing higher-level skills as outlined in the CCSS; and continual and timely monitoring of student achievement and progress via through-course and summative assessments (Bennett, Kane, & Bridgeman, 2011; Pruitt, 2011).

In addition to improving the usefulness of test scores, the PARCC assessment was designed to improve accessibility for both students with disabilities and English learners. This was accomplished through both item development following universal design principles (Center for Applied Special Technologies, 2011) and through the addition of numerous testing tools and technology-based accommodations into the computer-based assessment platform. While most features and accommodations were universally supported by member states, audio presentation of test content in the English Language Arts (ELA)/literacy assessment was one for which member states required further evidence to evaluate. Audio presentation of text can be useful to students who are unable to access test content due to visual disabilities (i.e., students who are blind or visually impaired) or due to learning disabilities (e.g., difficulties with decoding words in the process of reading). The text-to-speech (TTS) accommodation provides a delivery mechanism for audio presentation that removes such obstacles during administration as the need for a teacher or proctor to read aloud materials in a separate setting so as not to distract other students taking the test in standard conditions and to avoid providing the accommodation to students who do not require it. Further, the digital rendering of audio via TTS enables the student to control aspects of text delivery such as what is read and, depending on the platform, the voice

and rate of delivery that are used. However, there is a concern that assessing reading in the presence of an audio accommodation may change what is actually being measured by the assessment. The purpose of this study is to summarize and discuss current research on audio presentation via the TTS accommodation as it relates to the validity of the PARCC assessment as a predictor of college readiness. This vein of research is of particular importance to the field because there has not been similar research to date in the context of CCSS-aligned state assessments.

Predictive validity is one form of validity evidence for an assessment that associates that test's score with the criterion it is intended to predict, or to a proxy for that criterion. For example, college admissions tests typically claim to be predictive of student performance in college. Therefore, predictive validity studies for college admissions tests often focus on the relationship of the admissions test score to criteria such as first year (freshman) college GPA as a proxy for academic success in college. The PARCC assessment system claims relate scores on those assessments to performance level descriptors (PLDs), which "describe in broad terms the knowledge, skills, and practices students performing at a given performance level are able to demonstrate at any grade level" (PARCC, 2012). PLDs include claims about what inferences can be made about students at each performance level, and the scores associated with the PLDs then serve as a predictor of an outcome such as performance in an academic setting. From a fairness standpoint, validity aspects such as the interpretability and comparability of scores taken under different conditions and by different subgroups must be evaluated (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). One way to examine the predictive validity of test scores from tests taken with an accommodation is to compare the relationships of scores of those taken with and without the accommodation to the PLDs.

This report describes the design, recruiting process, experimental administration, and analyzes results from a research study intended to examine predictive validity of the PARCC test scores taken with and without TTS accommodations. In the course of the study, challenges during recruitment led to sample sizes too small to perform the quantitative analyses that were planned. Therefore, this report provides (a) an overview of the PARCC PLDs, (b) a summary of prior literature on audio presentation (read aloud), and (c) a summary of a study recently completed that examines the predictive validity of the PARCC test scores taken with and without TTS accommodations. We report sample characteristics and summary statistics of PARCC assessment performance and college course grades. Additionally, we provide lessons learned about the challenges that led to study limitations as well as about study design features that were promising or beneficial in order to aid researchers in designing future studies to contribute to this important research area.

Performance Level Descriptors

In 2012, the PARCC College- and Career-Ready Determination Policy in English Language Arts/literacy and Mathematics & PLDs were adopted by the PARCC Governing Board and Advisory Committee on College Readiness (PARCC, 2012). The document was updated in 2013 and revised in 2015. Since the current study is particularly focused on TTS for reading comprehension, this report focuses on the PLDs for the ELA/literacy assessments. The first of the two major components of this document define College and Career Readiness (CCR) Determination, a designation applied to students based on test performance that indicates the student has provided evidence of CCR. The Determination is described as follows:

> Students who earn a College- and Career-Ready Determination in ELA/literacy will have demonstrated the academic knowledge, skills and practices necessary to enter directly into and succeed in entry-level, credit-bearing courses in College English Composition, Literature, and technical courses requiring college-level reading and writing. (PARCC, 2012; p. 2)

This Determination requires that students achieve at least a Level 4 on the PARCC ELA/literacy assessment taken in grade 11 and provides postsecondary institutions with information that would allow them to exempt these students from taking remedial courses. It is noted carefully in the document that a Determination does not represent evidence of the full spectrum of skills and knowledge that may be required for CCR; further, a Determination should not be used as the basis for postsecondary admissions or for decisions about exemptions from more advanced coursework.

The second major component of the PARCC College- and Career-Ready Determination Policy in English Language Arts/literacy and Mathematics & PLDs is a set of policy-level PLDs. The PLDs describe what a student achieving each performance level should know and be able to do. A student is assigned to one of five performance levels based on their PARCC score. The thresholds for the performance levels were the result of a performance level setting process, which involved content experts reviewing test items and making judgments about how students at each performance level would perform on the items. Therefore, a performance level designation (e.g., Level 4) represents a categorization that distinguishes performance with respect to implications for the student (i.e., policy claims) as well as with respect to broad statements of knowledge, skills, and practices (i.e., general content claims).

Taken together, these two components — policy claims and general content claims — determine the framework from which claims can be made about the relationship between performance on the grade 11 PARCC ELA/literacy assessment and performance in entry-level, credit bearing English courses. PARCC's grade 11 ELA/literacy assessment claim is stated as follows:

> Students who earn a College- and Career-Ready Determination by performing at level 4 in ELA/literacy and enroll in College English Composition, Literature, and

> technical courses requiring college-level reading and writing have approximately a 0.75 probability of earning college credit by attaining at least a grade of C or its equivalent in those courses. (PARCC, 2012; p. 4)

This claim is the criterion for which PARCC commissioned the current study. Because there is an expected relationship of PARCC scores in grade 11 to the level of performance a student demonstrates in early college, a predictive validity study was requested. Predictive validity is a key part of the validity argument that supports the claims made about test score uses and interpretations (AERA, APA, & NCME, 2014). Because PARCC scores are intended to measure college and career readiness for all students taking the assessments, it is essential that the predictive value of PARCC scores applies to students taking either the accommodated or the nonaccommodated test form.

Literature Review on Audio Presentation Accommodation

What is an Accommodation?

Cortiella (2005) classified accommodations into four categories: presentation, response, timing and scheduling, and setting. Presentation accommodations include those in which the test material is conveyed to the student in a different format, such as via large print for students with visual impairments, or talking calculators for students who are blind. Audio accommodations such as human reader, audio tapes, and TTS fall into this category. Response accommodations are those by which the student may respond to test items in a different manner, such as by dictating responses to a scribe or responding in a test booklet instead of on an answer sheet. Timing and scheduling accommodations involve the provision of extra time, additional breaks, or splitting testing sessions across days or time periods. These accommodations are often invoked for students who suffer from fatigue or who must use accommodations that require extra time (e.g., large print may take more time for the student to scan the material). Setting accommodations include changes to the testing venue, such as allowing students to test in a separate room (e.g., to reduce distractions for the student, or to reduce distraction of other students when the student is having the test read aloud by the teacher), or allowing the student to test at home or in another location.

While all of these test changes can be classified generally under the umbrella of accommodations, they can have quite different influences on the validity of resulting test scores and interpretations. Therefore, accommodation policies are designed to provide accessibility while maintaining validity and may differ greatly depending on testing environments and goals. The different policies will lead to different consequences with respect to test scores. For example, test scores taken under conditions that alter the construct being measured may be flagged so that recipients of score reports interpret results with that fact in mind, may not be aggregated with test scores from tests taken under standard conditions because of a lack of

comparability, or may not be reported at all. In particular, audio accommodations on ELA assessments are controversial due to the potential change to the construct; and states have long had varying policies with respect to those test changes. Thurlow and Larson (2011) distinguish between state policies in which an accommodation is allowed, prohibited, allowed in certain circumstances, or allowed with implications for scoring. Of the states reporting policies, 37 states prohibited the reading aloud of passages, 15 states allowed read aloud of passages but included consequences with respect to scoring and reporting, and three states fully allowed the accommodation. It is important to note that the material that is being read aloud has a policy impact. Reading aloud directions is more universally allowed without consequences, and reading aloud test questions was prohibited by approximately 42 percent of policies as compared to approximately 48 percent of policies for reading aloud passages.

PARCC's policy, designed to be common across member states, distinguishes between three tiers of test changes: accessibility features for all students, accessibility features that are identified in advance, and accommodations (PARCC, 2016). The goal of allowing these test supports is to improve fairness and accessibility of the assessments for all students by ameliorating any obstacles to accessing test content that are unrelated to what is intended to be measured. Students must have a Student Registration/Personal Needs Profile (SR/PNP) that indicates which test changes are appropriate. Features available to all students on demand through the online platform include answer masking, adjustment of color contrast, magnification, and a pop-up glossary defining selected terms. Accessibility features that are identified in advance are also referred to as administrative conditions and mainly consist of timing and setting changes such as small group testing or testing at a different time of day. While they are available to all students, the principal or test coordinator must authorize such changes. Finally, accommodations are available on a restricted basis for students with disabilities and English learners. Accommodations include the use of assistive technology and screen readers, alternative formats of the test such as large print or braille, and American Sign Language (ASL) delivery for the ELA/literacy assessments. The PARCC Accessibility Features and Accommodations Manual states that accommodations should "(a) Provide equitable access during instruction and assessments; (b) Mitigate the effect of a student's disability; (c) Not reduce learning or performance expectations; (d) Not change the construct being assessed; and (e) Not compromise the integrity or validity of the assessment" (p. 20). Because accommodations represent more substantial changes to the administration of the test and may, as noted previously, change what is being measured, accommodations are allowed only for students who have significant need of them. The number of students receiving accommodations is minimized because the tests are designed from the development stage to be accessible, and the more widely available accessibility features are often adequate to provide any additional support that is required. As is noted in the PARCC Accessibility Features and Accommodations Manual, it is important that students using these accommodations in a testing situation be familiar with their use in the classroom; however, some accommodations allowed in the classroom may not be allowed on the

assessment because of the impact on measurement of the construct. When these test changes are invoked, the resulting scores may be excluded because of the lack of comparability with scores resulting from allowable conditions.

The use of TTS on an ELA/literacy assessment falls into the third category and is classified as an accommodation. From the PARCC Accessibility Features and Accommodations Manual:

> Purpose: The purpose of the embedded text-to-speech, ASL video, and Human Reader/Human Signer accommodation for the PARCC ELA/literacy assessment is to provide access to printed or written texts on the PARCC ELA/literacy assessments for a very small number of students with print-related disabilities who would otherwise be unable to participate in the assessment because their disability severely limits or prevents their ability to access printed text by decoding. This accommodation is not intended for students reading somewhat (i.e., only moderately) below grade level.
>
> Identification for SR/PNP: The student's SR/PNP must have text-to-speech, ASL Video, or Human Reader/Human Signer selected to activate the features on the platform. Once a student is placed into a session, the student will be assigned a form with embedded text-to-speech, or ASL Video.
>
> **Note:** There may be unintended consequences related to the use of this accommodation for some students. Review the adjacent Administration Guidelines carefully. PARCC will conduct additional research to provide PARCC states with data to substantiate the need for providing this level of access to a small number of students. (PARCC, 2016, p. 28)

It should be noted that TTS is allowed as a feature available to all students on mathematics assessments, when pre-identified in the SR/PNP, whereas TTS on ELA/literacy is an accommodation. This, again, is due to the construct intended to be measured. For ELA/literacy assessments, reading proficiency is typically a measurement target and is a product of the underlying components of decoding, fluency, and phonemic awareness, among others (Thurlow et al., 2009). Decoding is not typically considered to be a target of measurement on a mathematics assessment. Because constructs for different content areas may overlap, this distinction necessarily depends on how the construct is defined for a particular assessment.

What do Accommodations have to do with Test Validity?

Validity deals with interpretations of test scores. If the test score does not assess the intended construct(s), or if it assesses construct(s) not intended to be assessed (i.e., construct irrelevance), interpretation of a student's proficiency lacks validity. To take an extreme example, if a student who is blind received a low score on an ELA/literacy test, and the student took the test without

accommodations such as braille, it would be impossible to interpret that test score as indicative of low proficiency rather than as a result of an accessibility barrier. The scenario becomes more nebulous when we consider students with reading-based learning disabilities who take ELA/literacy tests. If the student has trouble decoding words but can comprehend text and reason through it when the text is known, then providing an accommodation such as TTS removes the barrier to demonstrate reading comprehension. However, the critical issue is how the construct being measured by the test is defined. If the construct of reading comprehension is defined as including decoding as a component, then providing TTS changes the construct being measured, leading to an unclear interpretation of the score with respect to proficiency and a lack of comparability between test scores of students taking TTS and students taking a nonaccommodated version of the same test. In both cases, the use of TTS may have an impact on the validity of the interpretation of the resulting test scores. Therefore, empirical research to examine and validate the claims of an assessment under different testing conditions is crucial.

Differential Boost Studies

One of the most commonly accepted ways to evaluate the appropriateness of an accommodation empirically is to compare whether and how much test scores increase under the accommodation for the target group versus for students who do not need the accommodation. The interaction hypothesis requires that an accommodation significantly increase scores only for students in the target group without a significant increase for the comparison group (Sireci, Scarpati, & Li, 2005). A modified approach requires only that there be a differential boost — that students in both groups may receive a significant score boost but that the score boost for students in the target group be significantly larger (see Cahalan-Laitusis, 2007). Herein we provide a brief overview of studies examining the appropriateness of audio accommodations from a differential boost perspective.[1]

Fletcher et al. (2006) randomly assigned accommodated or standard conditions to students with dyslexia or who did not have disabilities in grade 3. There were 44 students taking the accommodated version from each group and 47 students taking the standard version in each group. The accommodated condition included three test changes: (a) a timing/scheduling change, (b) reading proper nouns aloud to students, (c) reading stems and item options (not passages). The researchers used a three-level mixed model (i.e., students within schools within districts). The analysis of covariance (ANCOVA) included the Texas Assessment of Knowledge and Skills (TAKS) score as outcome; covariates were accommodation status, decoding status, vocabulary, and interactions, with random effects for school and district (6 districts, 48 schools). They also applied a standard (non-nested, fixed) loglinear model with whether the student met the state accountability standard as outcome and vocabulary, accommodation indicator, decoding group

[1] See http://nceo.umn.edu/docs/OnlinePubs/Report402/NCEOReport402.pdf for a recent, comprehensive summary of research on test accommodations.

indicator, and accommodation by group interaction. ANCOVA results showed a significant interaction of group and accommodation: $F(1, 55) = 12.04$, $p = 0.0007$, $d = 0.76$ (difference in effect size [ES] between students with and students without dyslexia; clustering not taken into account). The authors note that these are large effect sizes and are unadjusted for clustering (adjusting the pooled standard deviation would result in larger effect sizes but would be less comparable to other studies in the literature). The authors tried other random effect structures with no difference: Results satisfied the differential boost hypothesis regardless. The authors also did not find a relationship between extent of decoding difficulty and effect of accommodation in the dyslexic group. Looking at pass rates (loglinear model): In the dyslexic group, 41 percent using the accommodation passed, versus nine percent passing without the accommodation. The Wald chi-square of 12.82, $p < 0.0005$ was significant. The percentages translated into predicted odds of passing to be a seven-fold increase (0.695 vs 0.099, accommodated versus standard) for the dyslexic group. For students without disabilities, the pass rate was left statistically unchanged and in fact was a negative change (83 percent in standard versus 77 percent in accommodated).

Fletcher et al. (2009) applied similar test changes as in Fletcher et al. (2006) but in grade 7. However, the three conditions were standard, accommodated with read aloud as described in Fletcher et al. (2006) and administered in one day, and accommodated with read aloud as described in Fletcher et al. (2006) and administered in two days. Sample sizes for poor readers in those three conditions were, respectively, 56, 53, and 59 students; and the sample sizes for average readers were 66, 61, and 64 students. In this set of analyses, they did not account for clustering (17 schools in four districts). They used a linear model with TAKS score as outcome and decoding group, administration condition, and the interaction of the two as independent variables. A second linear model applied just to poor readers was used to evaluate the interactions of accommodation status with each ability covariate. The third type of analysis they did was to evaluate within-group chi-squares to look at how pass rates are affected by accommodation condition. The first linear model did not show a significant interaction of group and accommodation; therefore the accommodations do not meet the differential boost requirement. This interaction was not significant. The authors found for the second set of analyses that no ability covariates (decoding, fluency, critical reasoning, listening comprehension) affected the effects of the accommodations (all interactions with the accommodation condition were nonsignificant). The third set of analyses, executed within group, consisted of 3 x 2 chi-square tests of association between accommodation condition and pass rate. The results were significant for both groups — poor readers: chi-square $(2, N = 168) = 8.03$, $p = 0.018$; average readers: chi-square $(2, N = 191) = 7.07$, $p = 0.029$. In terms of pass rates, for the standard, one-day, two-day they were seven percent, 17 percent, 27 percent for poor readers and 68 percent, 84 percent, 85 percent for average readers.

Meloy, Deville, and Frisbie (2002) randomly assigned students within each grade (6, 7, and 8) who had no disability or had a reading-based learning disability to take the Iowa Tests of Basic

Skills (ITBS) under standard or read aloud conditions. Results reported were collapsed across grades, although each grade was administered a different, grade-appropriate level of the test (levels 12, 13, and 14 respectively). For the reading comprehension test, there were 98 students without disabilities who took the standard condition and 100 who took the read aloud condition, and there were 29 students with reading-based learning disabilities who took the test under standard conditions and 33 who took it with read aloud. Test performance was reported on the normal curve equivalent scale, which supports the collapsing of information across grades. The researchers performed a 2 x 2 ANOVA (disability group by accommodation status), $F(1, 260) = 37.54$, $p < 0.001$. There were significant main effects for disability group (e.g., students without disabilities scored uniformly higher) and accommodation status (e.g., scores were uniformly higher in the accommodated condition); but the interaction was not significant. Further, students with reading-based learning disabilities performed about 0.75 standard deviations (SD) better under read aloud than standard, whereas students without disabilities had a mean score increase of about 0.5 SD, leaving a differential gain of only 0.25 SD, far shy of the 1.0 SD criterion that has been proposed as a criterion for accommodation appropriateness (attributed to Fuchs, 1999).

Study Investigating Predictive Validity of PARCC ELA/Literacy Scores

Methodology

The current study was designed to examine the external validity of the read aloud/TTS accommodation offered by PARCC.[2] The TTS accommodation is intended for few students on the ELA/literacy assessments. In particular, only students who are blind and cannot or have not learned braille, or students who have a learning disability that limits their ability to decode text, are to have this accommodation — which is indicated in their individualized education plans (IEP). This is because the use of TTS and other audio accommodations has been criticized for altering the construct being measured on an ELA test (see Laitusis & Cook, 2008, for a summary of this debate).

Because the PARCC test is administered with this accommodation to some students, however, it is important to evaluate the validity of the assessment and interpretation of results under accommodated and nonaccommodated conditions. The current study was designed to compare predictive validity for:

A. general education students without an accommodation

B. students who need the TTS accommodation and are given this accommodation

[2] See https://parcc.pearson.com/tutorial/ for links to tutorials for the TTS accommodation and a guide to its functionality. The Maryland Assessments website also describes the PARCC TTS in more detail: https://marylandassessments.org/2014/03/24/text-to-speech-info-for-parcc-field-tests/.

C. students who need the TTS accommodation but are not given this accommodation

In addition, a group of general education students (i.e., Group D) was assigned to use the TTS accommodation in order to provide an evaluation of differential score boost between the groups. Such a differential boost would provide evidence supporting the appropriateness of the accommodation for the target group. The addition of Group D also took advantage of the benefit of easier recruitment of those students, providing universities with additional potential participants and additional income, which enhanced the incentive for institutions to participate. The study was designed to provide evidence to determine if the TTS accommodation removes construct-irrelevant barriers that prevent students from demonstrating their college readiness as measured by the PARCC grade 11 ELA/literacy assessment. The research questions to be investigated were:

1. Does performance on PARCC assessments predict first-year college students' course performance for Groups A, B, C, and D, as described previously?

2. How does the strength of the relations between performance on PARCC assessments and freshman course performance compare for the four groups?

3. How does performance on the grade 11 ELA/literacy assessment differ between students who do and do not require TTS in terms of score boost (accommodated minus nonaccommodated in each group)?

The aforementioned research questions can be evaluated under several data collection and analysis designs. Due to the need to test Groups C and D (students requiring the accommodation who will not receive it, and students who do not require the accommodation who will receive it) for the purposes of this study, an experimental design was required.

The rationale for using experimental design rather than analyzing extant operational data is twofold: First, students using accommodations may use bundled accommodations which would make disentangling the effects of TTS challenging; further, when testing operationally, students have the right to be tested using appropriate conditions for them, making random assignment of potentially impactful conditions impossible. Therefore, the current study used an independent-groups (alternatively, between-subjects) design, which is common for predictive validity studies. The design required random assignment of students needing the accommodation to either Group B or Group C and random assignment of general population students to either Group A or Group D. The outline of the design components follows.

High-level Design. The design consisted of the four groups of college freshmen taking the PARCC ELA Grade 11 Performance-Based Assessment (PBA) and End-of-Year Assessment (EOY) at agreed-upon university locations in fall 2015. We planned to analyze the score boost (i.e., accommodated minus nonaccommodated score) by group in order to determine whether

evidence of a differential boost exists. We also planned to use these scores to examine the predictive validity of the PARCC ELA Grade 11 test with respect to grades in entry-level credit-bearing English courses and/or first-semester and first-year grade-point averages (FSGPA, FYGPA). The study design and materials were submitted to the Committee for Prior Review of Research at ETS and through the institutional review board process at participating colleges that required that approval, and they were also reviewed by PARCC's Accessibility, Accommodations, and Fairness working group.

Participants. We targeted recruitment to first-year college students who attended high school in a PARCC state and graduated from high school in 2015, were enrolled in entry-level credit-bearing English courses, and who are either native English speakers or had reached at least the Intermediate or Developing level of English language proficiency. In the first stage of sampling, we recruited colleges and universities that were diverse with respect to selectivity and geographical location across PARCC states. We tried to include large public universities, private colleges, and community colleges so that the sample was representative of a distribution of "college ready" students. In an attempt to obtain adequate sample sizes, we targeted some institutions that have specialized programs for students with disabilities along with colleges and universities with typical services for students with disabilities. For example, we targeted the University of Arizona's Strategic Alternative Learning Techniques (SALT) Center — which services 500 students per year and had a recruitment presence in four PARCC states at the time of recruitment (i.e., Arizona, Illinois, New Jersey, and New York) — as a school that has specialized programs for students with disabilities. Because obtaining sample sizes proved to be prohibitively difficult when recruiting from PARCC states alone, we proposed including students from former PARCC states and students from other states that have implemented the CCSS, but we eventually opened recruitment to students from all states. Given that our sample consisted of students graduating from high school in 2015, it was not possible to guarantee that they had received CCSS-aligned instruction.

We anticipated the need to recruit up to 20 colleges and universities to participate in this study. We recruited schools by contacting both the Office of Institutional Research and the Disability Support Services office, or similar. We narrowed our focus to include only students who have indicated eligibility for the TTS accommodation through an IEP or 504 plan in their high school instruction and/or high school standardized testing. We attempted to recruit 100 students in each of Groups A through D. Interested students were directed to a web survey that collected background information for screening purposes, such as the state in which the student attended grade 12; year of high school graduation; birth date; native language; audio accommodation eligibility; use of audio tools and accommodations inside the classroom, outside the classroom, and on tests; and diagnosed disabilities. See Appendix A for a copy of the screening survey. When students expressed an interest in participating, they were registered for the testing platform and randomly assigned TTS or no TTS within disability group — a requirement for the test

administration. Students who required accommodations that would not be provided and that were not part of the study focus were not eligible to participate.

Materials. The PARCC ELA Grade 11 PBA and EOY test were used for the study. One ELA Grade 11 PBA form and one EOY form included TTS as an embedded accommodation. In addition, an alternate form containing the same items did not have TTS available. These test forms were operational test forms from an earlier testing window, and both test forms contained the same items. The EOY form has 22 operational items worth a total of 44 points. The PBA form has 23 items worth a total of 93 points. Therefore, there are 137 total raw score points available on the form. Both PBA and EOY components were administered in separate sessions, mimicking the operational administration as closely as possible. With approximately three to four hours of testing for PBA and two to three hours for EOY, the total amount of time required per participant was approximately six hours across two testing sessions.

Although the PBA is typically completed earlier in the school year than the EOY components for the operational PARCC, we administered EOY first so that we could obtain a complete set of EOY data in case of attrition. At the end of the EOY session, students completed a brief post-test survey that asked about high school cumulative GPA, ELA course grades, admissions test scores, broad college major (i.e., Natural Sciences, Engineering, Social and Behavioral Sciences, Arts and Humanities, Education, Business, and Other Fields), and interest in future research participation. See Appendix B for a copy of the survey. All proctors viewed a webinar created by Pearson that described the platform and were provided with tutorial materials and an administration guide for the PARCC operational assessment.

Payments. Given the relatively small sample sizes available for this study and little intrinsic incentive for schools or students to participate, we offered honoraria for both institutions and student participants. We provided student participants with $150.00 for participating in both the PBA and EOY testing sessions, with payment dependent on completion of both sessions. We compensated schools at the rate of $100.00 per student under the assumption that the Registrar's office would cooperate with the data collection coordinator in providing student fall 2015 and spring 2016 grades and course categories.

Planned Analyses

As was noted previously, recruitment obstacles led to sample size deficiencies that could not be overcome in order to complete our planned analyses. However, we report our full analysis plan here with the goal of providing future researchers a potential pathway toward evaluating the validity and fairness of accommodations in similar contexts.

Predictive Validity Analysis. As previously mentioned, PARCC's grade 11 ELA/literacy assessment claim is stated as follows:

> Students who earn a College- and Career-Ready Determination by performing at level 4 in ELA/literacy and enroll in College English Composition, Literature, and technical courses requiring college-level reading and writing have approximately a 0.75 probability of earning college credit by attaining at least a grade of C or its equivalent in those courses. (PARCC, 2012; p. 4)

In order to evaluate this claim, we planned to examine mean PARCC grade 11 ELA scores for the four recruited groups in the following grade categories based on their entry-level ELA course(s): did not qualify (remediated), dropped out, F, D, C, B, or A.

Differential Validity Analysis. We planned to obtain the correlations of the PARCC ELA Grade 11 test with the criterion (ELA grade, FSGPA, FYGPA) by group, followed by typical statistical tests used to compare the strengths of the correlations — after using Fisher's *r*-to-*z* transformation. The correlations, or validity coefficients, indicate the prediction from the PARCC score of GPA or grade as the case may be.

Differential Prediction Analysis. We then planned to use the combined-group prediction equation to obtain residuals from that line for each group that indicate under- or over-prediction. This would provide evidence of whether the relationship of PARCC as a predictor of the criterion is similar or systematically different for the groups. Since PARCC is a college readiness assessment, the use of first-year college GPA as a criterion of college readiness was selected in addition to the more specific criterion of ELA grade. Most predictive validity studies on testing accommodations for the SAT or ACT have used FYGPA as the criterion.

Differential Boost Analysis. This type of analysis requires either repeated-measures data (i.e., each student in each group takes the test under both accommodated and nonaccommodated conditions) or independent-groups design with random assignment, which occurred in this study. Score boost is compared across groups, and whether the group requiring accommodations received a significantly greater score boost than the general population group provides further evidence of whether the accommodation is appropriate in this context.

For the differential boost analysis, we refer to two independent, concurrent meta-analyses on read aloud for students with disabilities (Li, 2014 and Buzick & Stone, 2014) and a critical research synthesis that further describes many of those studies in more detail (Laitusis, Buzick, Stone, Hansen, & Hakkinen, 2012). There has been only one experimentally designed study of TTS on ELA/literacy (Olson & Dirir, 2010). The Connecticut study that is part of this report used a repeated-measures design with approximately 200 students with disabilities eligible for read aloud and 200 students without disabilities, all in grade 7. The two tests consisted of parallel halves of a full test form, with 20 multiple-choice items each. Although there were no significant differences found for either group between conditions, the accommodation condition order was

not counterbalanced (standard was always administered first), confounding any possible inferences.

Examining all experimental read-aloud studies in Figure 1 of Buzick and Stone (2014), it is clear that effect sizes for read aloud studies on ELA/literacy range from 0.29 to 1.12, with an average of 0.62 for students with disabilities. For students without disabilities, the effects sizes range from 0.03 to 0.70, with an average of 0.25. The confidence intervals for the differences in boost between groups, shown in Figure 2 of that article, are positive and did not cross 0 for five out of eight studies evaluated. Our design consists of independent groups. The three independent-groups studies that fit the criteria for the meta-analysis (marked with "~" after their abbreviated citations in the Figures) had study sample sizes of 47 and 44, standard and accommodated, respectively; and 56 and 53, and 29 and 33 for the groups of students with disabilities. Note that these studies also produced two of the largest effect sizes for students with disabilities.

The target sample sizes, 100 participants per subgroup, should have provided adequate power for the analyses. In fact, given the large effect sizes reported in the audio accommodations literature, there is some indication that sample sizes as low as 25 per experimental group would have sufficed (Buzick & Stone, 2014). As mentioned previously, differential boost analyses use ANOVA methods. Actual power analyses would require some knowledge of the expected means and standard deviations of scores for the groups under different conditions; however, we do not have this information, nor do we have effect sizes for digital read aloud (TTS) in this context. While it might be useful to infer possible effect size projections from similar research, there were no studies of read aloud on ELA found at the high school level (Buzick & Stone, 2014).

One study did use TTS on state test items at grade 7. However, the study design was not counterbalanced, so no conclusions can be drawn about the effect of the accommodation. Further, there has been little research on performance-based assessments with respect to the read aloud accommodation, making it difficult to project possible effects in that format. Differential validity studies generally use correlation, whereas differential prediction studies use regression. These approaches typically include analysis by institution with subsequent pooling of results across institutions. This places requirements on sample sizes at both the institution and participant levels. While correlations can be computed for groups with as few as five individuals in each institution and some rules of thumb suggest seven, there may be resulting instability in most cases. Therefore, it may be optimal with a fixed total sample size to recruit from as few institutions as possible and to maximize the number of individuals per institution (see, e.g., Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008, in which a minimum subgroup size per institution of 15 was chosen for this reason). Similar sample sizes should be more than adequate for differential prediction, in which regression equations would be computed on the total number of students within school, assuming few predictors (e.g., composite score on the research PARCC ELA form) that capitalize on information across schools if needed to meet sample size requirements.

Results

A total of approximately 370 institutions in the United States were contacted through personalized e-mails to administrators and disability services staff with follow up by telephone or e-mail as requested by the institution. In June 2015, researchers also presented information about the study to the PARCC Higher Education Leadership Team, which includes members who are highly placed administrators within institutions and institutional systems. A presentation was made at the Association for Higher Education and Disability (AHEAD) in July 2015 to gain feedback from audience members as well as to seek possible institutional partners, and information about the study was shared to institutional service providers via disability services e-mail lists. Researchers also reached out to friends and colleagues at institutions in order to spark recruitment. However, these approaches did not yield adequate sample sizes due to challenges described below. Due to recruitment deficiencies at the institutional level, a second wave of recruitment took place in which students local to ETS were tested by ETS staff. In this second wave, students were still paid $150.00; however, they received $100.00 after both testing sessions and $50.00 after all fall 2015 and spring 2016 grades were sent to ETS because they were self-reported. In all, a diverse group of 15 institutions were represented.

Table 1 describes the planned and actual sample sizes obtained. Clearly, the number of students actually recruited and randomly assigned to each group is quite a bit smaller than the original study design called for, resulting in a change to examining descriptive sample statistics rather than evaluating differential boost and a plan to collapse grade categories for the predictive analysis.

Table 1. Goal and Actual Sample Sizes for Text-to-Speech (TTS) Validation Study.

Group	TTS Provided	Goal N (Actual N)	Session 1	Session 2
A. General Education	No	100 (19)	EOY	PBA
B. SWD who require TTS	No	100 (8)	EOY	PBA
C. SWD who require TTS	Yes	100 (5)	EOY	PBA
D. General Education	Yes	100 (14)	EOY	PBA

Note that the assignment within disability group to use or not use TTS is unbalanced. This is due to the attrition between recruitment and testing by numerous students who agreed to participate but then did not schedule sessions.

Tables 2 through 8 describe results of the post-test survey. The sample was made up of students who reported being academically strong overall and in English courses, with correspondingly strong admissions test scores and with a spread of majors predominantly in Natural Sciences and Business.

Table 2. Responses to Post-test Survey Question 1 Regarding Cumulative High School Grade Point Average.

Indicate the range of your cumulative grade point average for all academic subjects in high school.	A (90-100)	B (80-89)	C (70-79)	D (Below C)	Missing
	19	21	3	0	3

Table 3. Responses to Post-test Survey Question 2 Regarding High School English Language Arts Grades.

What were your grades in your English Language Arts (ELA) courses in high school?	A Mostly A's and B's	B Mostly B's	C Mostly B's and C's	D Mostly C's	E Mostly below C's	Missing
	29	6	8	0	0	3

Table 4. Responses to Post-test Survey Question 3 Regarding Whether Students Took the ACT Admissions Test.

Did you take the ACT prior to college admissions?	Yes	No	Missing
	16	27	3

Table 5. Responses to Post-test Survey Question 3 Regarding Self-Reported ACT Component and Composite Scores.

Average scores for those taking the ACT:	Composite (0-36)	English (0-36)	Reading (0-36)	Writing (0-36)
	26.64	27.58	26.75	24.33

Table 6. Responses to Post-test Survey Question 4 Regarding Whether Students Took the SAT Admissions Test.

Did you take the SAT prior to college admissions?	Yes	No	Missing
	37	6	3

Table 7. Responses to Post-test Survey Question 4 Regarding Self-Reported SAT Component Scores.

Average scores for those taking the SAT:	Critical reading (200-800)	Writing (200-800)	Essay subscore (2-12)
	598.18	632.81	9.00

Table 8. Responses to Post-test Survey Question 5 Regarding College Major at the Time of Study Participation.

What is your college major?	Natural Sciences	Engineering	Social and Behavioral Sciences	Arts and Humanities	Education	Business	Other Fields	Missing
	8	0	3	2	2	9	19	3

Only a few participants experienced technical issues during their testing sessions. For example, during the first two testing sessions the proctor noticed there was some difficulty initializing the testing platform. In some cases the browser did not recognize that Java had been installed on the computer, or the pop-up blocker would block the window the testing system would initialize. To address these issues, Java was made active in the plugins manager of the browser and the TestNav™ delivery site[3] was added to the exceptions list of the pop-up blocker. Once these changes were made to the browser's settings, both issues were resolved.

[3] http://www.pearsonassessments.com/largescaleassessment/products-services/testnav.html

Another difficulty experienced involved participants being exited out of their testing sessions because a program running in the background of the computer shifted the focus away from the testing window. This was almost always caused by a scheduled maintenance or upgrade asking for permission to perform a task. To resolve this issue, the taskbar was checked at the beginning of each day of testing to make sure these tasks were either rescheduled for another day or performed before testing sessions were to begin.

Most participants were able to finish each session in roughly two hours or slightly less; however some students taking the TTS form took longer than anticipated and in several cases did not finish the PBA portion of the test under that condition. Because the standard and TTS forms consisted of the same items, raw scores could be compared between forms. Table 9 describes the summary statistics of raw scores for students participating in the study. From the table, several patterns emerge. Regardless of disability group, students taking the standard form of the test performed better on average than those taking the TTS form for the EOY portion of the assessment; while sample sizes are smaller, students taking the TTS form of the test performed better on average on the PBA portion when taking the TTS form. Additionally, the students who do not need TTS performed better on average than did students who do need TTS. Out of 137 possible raw points, scores for students who do not require TTS ranged from 0 – 118 on the standard form and 1 – 116 on the TTS form. Students who do require TTS had scores ranging from 22 – 98 on the standard form and 0 – 110 on the TTS form.

Table 9. Summary Statistics for Students Participating in the TTS Validity Study.

Test		Assigned TTS			Assigned Standard		
(Max. score)		*N*	Mean	SD	*N*	Mean	SD
EOY (44)	Needs TTS	5	20.80	14.32	8	24.38	9.15
	Does not need TTS	14	23.21	8.56	19	25.26	9.43
PBA (93)	Needs TTS	2	*	*	8	45.25	20.46
	Does not need TTS	12	55.08	15.89	18	50.83	18.96
Total (137)	Needs TTS	5	51.00	52.55	8	69.63	28.02
	Does not need TTS	14	70.43	32.49	19	73.42	30.36

Note: The sample size was too small for means and standard deviations where indicated by asterisk (*). These students received high scores on the PBA. Although EOY and PBA are not reported separately in an operational context, the effect of TTS may differ between the components; therefore statistics are given separately and for the total test.

Due to the low sample sizes, we were not able to evaluate the correspondence of PARCC score levels with individual grades (i.e., A, B, C, D, F, and remediated) and considered combining grades into categories of C and above and below C in order to directly examine the claim that students receiving a Level 4 or above on the PARCC assessment would have a 0.75 probability of earning a C or above in their entry-level, credit-bearing college English courses. However, when grades arrived, we realized that all but one participant had received a grade of C or above. Therefore, this simpler analysis was also impossible to perform. This grade distribution is in line with the self-reported high school ELA grades conveyed by students in the post-test survey, as described previously.

Limitations and Challenges

Many of the implementation challenges occurred at the time of participant recruitment, leading to limitations with respect to the planned analyses.

Recruitment Timeline and Potential Sample. Recruitment was initially planned to start in the spring semester of 2015; however, permissions required from PARCC prior to the start of recruitment arrived later than planned. By that time, some administrators who would be needed to make the decision to allow their institution to participate in the study were on vacation. At other institutions, they were undergoing turnover or replacement of those key personnel or had done so recently, making it infeasible to make the commitment needed to participate. Another aspect related to time has to do with the window of opportunity which institutions had to identify students who would be eligible for TTS and recruit them. Some students did not self-disclose their disability status and accommodation or accessibility needs until weeks into their first semester. This severely limited the ability of institutions to identify and reach out to these students in time to get them set up to test and to have their two testing sessions. Institutions also reported not having enough students with disabilities to participate.

Infrastructure. Institutions interested in participating had to provide several key elements in order to administer the test, including staff, space, and possibly computers. Staff would have the responsibility for identifying and/or coordinating registration of students to participate. Staff were also required to proctor and monitor a secure test administration for each participant, working with ETS and Pearson to answer student questions about the test administration and study. These staff members would have to complete Pearson-provided training in order to understand the Pearson TestNav 8™ testing platform (e.g., how to log students in or pause a test). Staff were asked to obtain student grade reports at the end of the fall 2015 and spring 2016 semesters. These requirements proved burdensome to institutions that already had an abundance of work and a dearth of staff, particularly disability services offices that also had to administer other testing to students throughout the semester. Students would also have to test in groups in a secure fashion or would have to test individually, requiring adequate physical space and technological resources.

In addition to these logistical obstacles, invited institutions declined to participate for a number of policy or perception reasons. Some institutions expressed a hesitance to participate because of political issues surrounding standardized testing or the CCSS. Further, numerous institutions reported already participating in other PARCC research (e.g., institutions in Massachusetts taking part in the PARCC/Massachusetts Comprehensive Assessment System [MCAS] validity study that had also recruited college freshmen: Nichols-Barrer, Place, Dillon, & Gill, 2015) or already participating in other assessments. Several institutions cited policies preventing research studies involving their students. A challenging aspect of recruitment is that institutions cannot be contacted until a human subjects review is complete and the study is approved. Therefore, it is important to finalize documents and procedures as early in the study as possible in order to get that review underway. However, when institutions are then contacted, it is not uncommon to discover that, for example, the institution has its own human subjects or institutional review board procedures which will push recruitment at the student level out further in the overall timeline. This was the case for one institution that was very interested in participating in the study but would not have been able to get the review done in time. Maintaining a database of institutional features may help to ameliorate some of these issues in the future. However, it is difficult to build a database comprehensive enough to capture all the features that would be relevant to any conceivable study.

As was previously noted, even with intentional focus on randomly assigning students within disability group to test with or without TTS, it is clear that the assignment was unbalanced with respect to sample size. Because of the registration mechanism, this may have been unavoidable. Students were registered in the delivery system as we received their names as willing participants, and they had to be assigned testing conditions at that time. However, the mechanism itself should have reduced or eliminated the possibility of a systematic aspect to that imbalance. Students also self-reported their disability status and accommodation eligibility, which out of necessity raises the question of the accuracy of all participant information. An additional issue is with respect to self-selection of the students to participate and the resulting lack of diversity of academic strength across the sample, preventing the examination of the predictive claim. Even with much larger sample sizes, it is possible that the selection bias or grade inflation could have precluded the evaluation of the predictive claim. One final aspect to take into account when examining effects of accommodations is the need to ensure that students in experimental groups taking an accommodation are familiar with the accommodation as implemented; otherwise, differential effects of the accommodation cannot be disentangled with certainty from differential exposure to and mastery of it.

Areas of Promise and Implications for Future Research and Practice

As designed, this would have been an important study because the research basis for PARCC's accommodation policy for the use of TTS is founded upon research studies with several key differences from the current context. First, there are few read aloud ELA studies focused on high

school or early college students, none of which counterbalanced accommodation modes or included students with learning disabilities (see Li, 2014 and Buzick & Stone, 2014). Second, none of the studies evaluating read aloud on ELA focused on Common Core assessments, which have been claimed to be more difficult than previous assessments;[4] on technology-enhanced items, with which the TTS may interact differently than with more straightforward item formats; and on a performance-based assessment, which requires the completion of longer writing tasks. Therefore, the research basis for the policy on this controversial accommodation is quite limited, and a full-scale study of this type would provide an opportunity to ensure the validity of the assessment results and interpretations when TTS is used on ELA/literacy by a target group of students who need the accommodation in these types of contexts.

Of further benefit is the ability to do this research on an operational PARCC test within the operational testing platform. It is crucial to try to obtain empirical evidence in as authentic a context as possible in order to make accurate inferences about the results. Therefore, the use of an operational test form with innovative item types, on the actual delivery platform, with the actual TTS mechanism, provides a substantial benefit over studies in which the testing agency is not involved with or does not contribute to the research study. Of further importance is the need to describe participants, materials, and platforms in adequate detail for readers and future researchers to fully understand results and implications and to combine results with other studies or to build upon the results shared. Finally, when generating random assignment to conditions, it would be optimal to be able to do so at the time of participation rather than at the time of recruitment, if logistically feasible, to prevent imbalances in the random assignment. When small sample sizes are obtained, as is too often the case in research regarding students with disabilities, it would also be worthwhile to investigate combining data across similar schools or using more robust models or nonparametric approaches.

While this study did not yield the conclusive empirical results that were the original goal, it is our hope that by sharing the details of our design and how the study was carried out, we can provide a basis for further research into this important accommodation and context. One aspect of interest for future research is to examine group performance at the item and item-type levels both overall and as it ties in to accommodation usage. We also hope that this report can shed light on the process by which testing agencies and states evaluate their accommodation policies in the interest of providing fair and valid assessments for all students.

References

[4] See, for example http://www.excelined.org/common-core-toolkit/old-standards-v-common-core-a-side-by-side-comparison-of-english-language-arts-2/, http://www.curriculumassociates.com/products/ready-most-challenging-common-core-standards-overview.aspx#, and http://www.usnews.com/news/articles/2016-02-22/common-core-tests-assess-student-achievement-differently.

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.

Bennett, R.E., Kane, M., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Princeton, NJ: Center for K-12 Assessment & Performance Management, Educational Testing Service.

Buzick, H. M., & Stone, E. A. (2014). A meta-analysis of research on the read aloud accommodation for K-12 students with disabilities. *Educational Measurement: Issues and Practice 33*(3), 17–30.

Cahalan-Laitusis, C. (2007). Validity and accommodations: A variety of approaches to accessible assessments. In C. Cahalan-Laitusis & L. L. Cook (Eds.), *Large scale assessment and accommodations: What works* (pp. 71-83). Washington, DC: Council for Exceptional Children.

Center for Applied Special Technologies (2011). *Universal design for learning guidelines version 2.0*. Wakefield, MA: Center for Applied Special Technologies.

Cortiella, C. (2005) *No Child Left Behind: Determining appropriate assessment accommodations for students with disabilities*. Retrieved from http://www.readingrockets.org/article/10938

Fletcher, J. M., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., & Vaughn, S. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children, 72,* 136–150.

Fletcher, J. M., Francis, D. J., O'Malley, K., Copeland, K., Mehta, P., Caldwell, C.J., & Vaughn, S. (2009). Effects of a bundled accommodations package on high-stakes testing for middle school students with reading disabilities. *Exceptional Children, 75*, 447–463.

Fuchs, L. S. (1999). *Curriculum-based measurement: Updates on its application in standards-based assessment systems*. Charlotte, NC: Council for Exceptional Children.

Laitusis, C., Buzick, H., Stone, E., Hansen, E. & Hakkinen, M. (2012). *Literature review of testing accommodations and accessibility tools*. Commissioned Report for the Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/08/Smarter-Balanced-Students-with-Disabilities-Literature-Review.pdf

Laitusis, C. C., & Cook, L. L. (2008). *Reading aloud a test of reading comprehension* (ETS Research Spotlight). Princeton, NJ: Educational Testing Service.

Li, H. (2014). The effects of read-aloud accommodations for students with and without disabilities: A meta-analysis. *Educational Measurement: Issues and Practice 33*(3), 3–16.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. (2008). *Differential validity and prediction of the SAT* (College Board Research Report No. 2008-4). New York, NY: The College Board.

Meloy, L. L., Deville, C., & Frisbie, D. A. (2002). The effect of a read aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education, 23*, 248–255.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010a). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. Washington, DC: Authors.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010b). *Common Core State Standards for Mathematics*. Washington, DC: Authors.

Nichols-Barrer, I., Place, K., Dillon, E., & Gill, B. (2015). *Predictive validity of MCAS and PARCC: Comparing 10th grade MCAS tests to PARCC Integrated Math II, Algebra II, and 10th grade English Language Arts tests*. Cambridge, MA: Mathematica Policy Research.

Olson, J. F., & Dirir, M. (2010). *Technical report for studies of the validity of test results for test accommodations—Establishing the validity of test accommodations and score interpretations for students with disabilities: A collaboration of state-based research.*

Partnership for Assessment of Readiness for College and Careers (2012). *PARCC college- and career-ready determination policy in English Language Arts/literacy and mathematics & policy-level performance level descriptors*. Retrieved from http://www.parcconline.org/files/79/College%20and%20Career%20Ready/92/PARCCCCRDPolicyandPLDsFINAL.pdf

Partnership for Assessment of Readiness for College and Careers (2016). *PARCC Accessibility Features and Accommodations Manual* 2016–2017 (5th ed). Washington, DC: PARCC Assessment Consortium.

Pruitt, S. (2011). Overview of the Partnership of Assessment of Readiness for College and Careers (PARCC). Retrieved from http://images.pearsonassessments.com/images/NES_Publications/2011_06Pruitt_PARCC.pdf

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457-490.

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments.* Minneapolis, MN: National Accessible Reading Assessment Projects.

Thurlow, M. L., & Larson, J. (2011). *Accommodations for state reading assessments: Policies across the nation.* Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Appendix A. Recruitment Survey

1. Did you graduate from high school in 2015?

a. Yes (proceed to question 2)

b. No (say thank you for your interest, you are not eligible to participate in this study. Would you like to be considered for future ETS research? [Yes/No/Maybe] If yes or maybe - Please provide your e-mail address so we can contact you about future research.)

2. In which state did you attend 12th grade? (Drop down box)

3. Are you currently enrolled in a college-level, credit-bearing English Language Arts class that teaches basic reading and writing? Some example course names are: Argument and Persuasion, Composition I, College Writing I, Freshman English I, Expository Writing, Rhetoric and Composition, Writing Rhetorically, Introduction to College Writing, Critical Writing and Reading, or Writing Course Sequence. (Yes/No)

4. What is your native language? (English/Other)

a. If student says English proceed to question 4.

b. If student chose Other, proceed to ELP question.

4. What is your current level of English language proficiency?

a. Proficient/Advanced/Reclassified

b. Intermediate/Developing

c. Emerging/Beginning

d. Unsure/Don't Know

If participants indicate Emerging/Beginning say thank you for your interest, you are not eligible to participate in this study. Would you like to be considered for future ETS research? (Yes/No/Maybe) If yes or maybe - Please provide your e-mail address so we can contact you about future research.

5. In high school, were you eligible to receive an audio accommodation (e.g., tests or classroom materials read aloud)? (Yes/No) If yes, proceed to question 6. If no, proceed to question 8.

6. Which of the following audio presentation tools or accommodations have you used in your high school English classes, high school standardized tests, and outside of the classroom? (Check all that apply)

	High School English Classes	High School Standardized Tests	Outside the Classroom
None			
Built in **text-to-speech** audio presentation tools such as: Apple VoiceOver, Windows Narrator, Speak Text for Microsoft Word, etc.			
Audio presentation **screen reader** software or program(s) such as: Jaws, Kurzweil, NVDA, Window-Eyes, or Balboaka Audiobooks or other pre-recorded audio			
Teacher or assistant (to read aloud material)			
Other (please specify) _____			

7. Do you use your read aloud tool with synchronized highlighting during instruction or while taking tests? (Yes/No/Sometimes - with open-ended box to explain)

	High School English Classes	High School Standardized Tests
Yes		
No		
Sometimes		

Explain:

8. Do you have any diagnosed disabilities? (Yes/No)

If yes, please select one or more of the following answers. Responses are a drop down box. Include the list of 13 federally identified disabilities allowing to select more than one, none, do

not wish to answer, or other disability or condition with an open-ended box to enter the disability or condition.

Appendix B. Post-test Survey.

Q1

Indicate the range of your cumulative grade point average for all academic subjects in high school.

- A. A (90-100)
- B. B (80-89)
- C. C (70-79)
- D. Below C

Q2

What were your grades in your English Language Arts (ELA) courses in high school?

- A. Mostly A's and B's
- B. Mostly B's
- C. Mostly B's and C's
- D. Mostly C's
- E. Mostly below C's

Q3

Part A

Did you take the ACT prior to college admission?

- A. Yes
- B. No

Part B

If you answered Yes to Part A, what is the highest score that you received in the following categories? (All scores range from 1 to 36)

Composite: [ ]

English: [ ]

Reading: [ ]

Writing (optional): [ ]

Q4

Part A

Did you take the SAT prior to college admission?

- A. Yes
- B. No

Part B

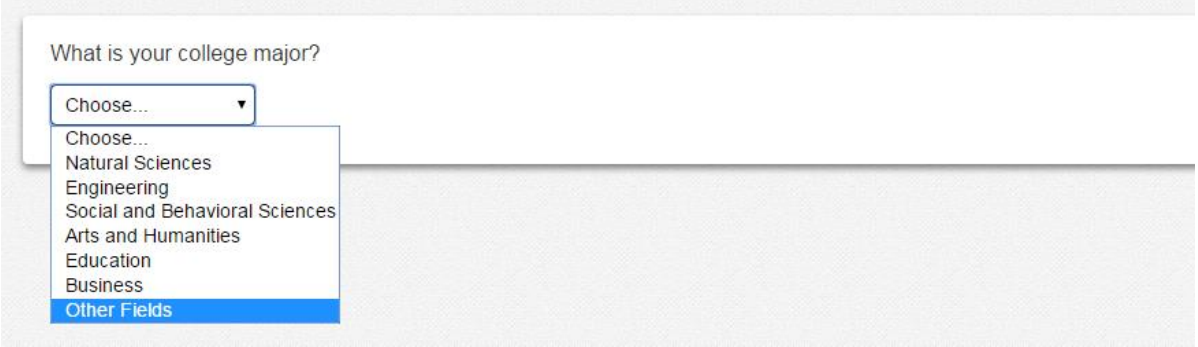
If you answered Yes to Part A, what is the highest score that you received in the following categories?

Critical reading (200-800): [ ]

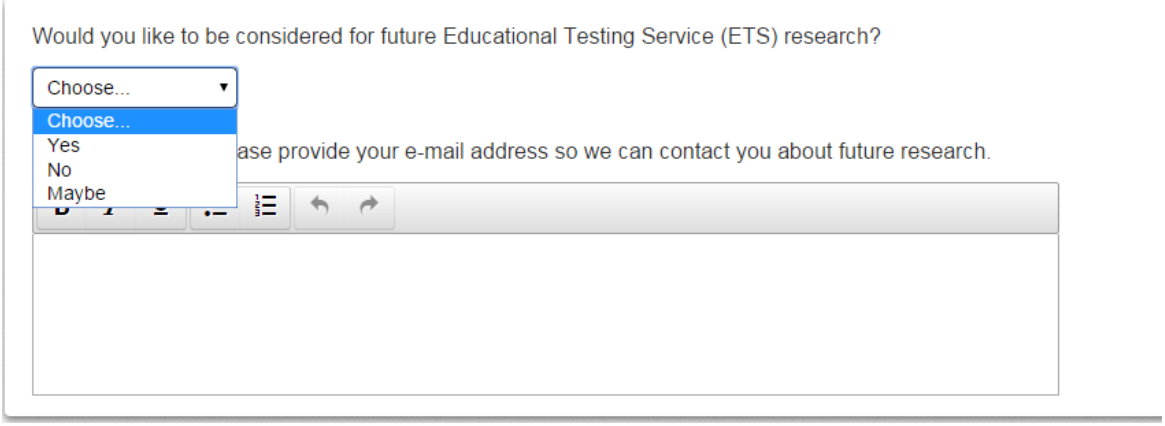
Writing (200-800): [ ]

Essay subscore (2-12): [ ]

Q5


What is your college major?
Choose...
Choose...
Natural Sciences
Engineering
Social and Behavioral Sciences
Arts and Humanities
Education
Business
Other Fields

Q6


Would you like to be considered for future Educational Testing Service (ETS) research?
Choose...
Choose...
Yes
No
Maybe
ase provide your e-mail address so we can contact you about future research.