Working Paper

2021-R2124

Score Gains and Validity Evidence for English Learners Testing with Supports on the ACT

JOANN L. MOORE, PHD, DONGMEI LI, PHD, AND YANG LU, PHD

ACT.org

Score Gains and Validity Evidence for English Learners Testing with Supports on the ACT

Joann L. Moore, PhD, Dongmei Li, PhD, and Yang Lu, PhD

Abstract

The purpose of this study was to provide validity evidence for English learners (ELs) taking the ACT® using testing supports. Three groups were compared to evaluate performance, score gains, and relationships between ACT scores and high school grades: ELs who took the ACT without testing supports and retested using testing supports, ELs who tested twice without supports, and non-ELs who tested twice without supports. ELs who retested with supports had the lowest ACT performance, followed by ELs who tested twice without supports, and non-ELs had the highest performance. Score gains were highest for ELs who retested with supports, especially in reading. Relationships between high school grades and ACT scores were higher on the second test attempt for ELs who retested with supports, particularly in reading and science. A second sample was used to evaluate psychometric properties of scores for ELs who tested with supports compared to ELs who tested without supports and non-ELs who tested without supports. Classification accuracy and consistency rates were generally high across groups. Some evidence of DIF was found between ELs who tested with supports and non-ELs, but about half favored the focal group and half favored the reference group; therefore, the impact on total scores is expected to be minimal. CSEM was virtually identical across groups, while score reliability and SEM were both lower for ELs who tested with supports compared to the other two groups. This study provides evidence that testing supports are providing a benefit to ELs and removing construct-irrelevant variance without conferring an unfair advantage. Future research is needed to compile additional validity evidence and examine the impact of the supports on predicting college performance.

Keywords: English learners, accommodations, testing supports, standardized testing, admissions testing

Score Gains and Validity Evidence for English Learners Testing with Supports on the ACT

Objectives

In 2017, ACT began providing testing supports (or accommodations) to eligible English learners (ELs) in the US taking the ACT® test, including translated test instructions, use of word-to-word bilingual dictionaries, small-group testing, and extra time. The goal of these supports is to remove construct-irrelevant variance and allow students with limited English proficiency to more accurately demonstrate their true achievement level. The purpose of this study is to investigate the effect of these supports on removing construct-irrelevant variance, thus resulting in ACT scores that more accurately reflect ELs' achievement level in the subject being assessed.

Theoretical Framework

The US Department of Education defines an EL as an individual between the ages of 3 and 21 who is enrolled or planning to enroll in elementary or secondary education; whose native language is not English; and whose lack of proficiency in reading, writing, speaking, or listening may deny them the ability to meet state academic standards; succeed in classrooms where instruction is in English; or be able to fully participate in society (US Department of Education, 2016). This is the definition of EL that is used by ACT and throughout this paper, although it should be noted that the procedures for identifying ELs (e.g., choice of English proficiency assessment, choice of cut scores) are determined at the district or state level.

Approximately five million students in the US are identified as English learners (ELs), and the percentage of ELs has increased from 9.2% in 2010 to 10.2% in 2018 (NCES, 2021). ELs are a diverse group of students, differing with respect to native language, time in the US, educational experiences, level of English proficiency, and many other factors. ELs are more likely to come from families with lower income and less parental education and to identify as a member of a traditionally marginalized racial/ethnic group, all of which are associated with lower levels of academic achievement (Abedi, 2002; Herman & Abedi, 2004). ELs tend to be exposed to fewer core academic courses in high school, graduate at lower rates, and enroll in college at lower rates than their English-proficient peers (Callahan & Shifrer, 2016; Johnson, 2019; Sugarman, 2019). ELs also tend to perform at a much lower level on standardized tests, in part due to lower proficiency in English, which can impact their ability to demonstrate their true achievement level (Abedi, 2002; Sugarman, 2019), although the factors listed above (demographics, school experiences) also likely play a substantial role in ELs' lower performance (Moore, in press).

Testing supports are modifications to a test or test administration conditions that are meant to reduce construct-irrelevant variance, allowing for more accurate measurement of a construct and resulting in scores that more accurately reflect students' knowledge, skills, and abilities. Testing supports are often provided for students with disabilities (e.g., providing a braille version of a test to students with visual impairments) and ELs (e.g., providing word-to-word bilingual dictionaries to native Spanish speakers). When assessing ELs in content areas other than English, their scores may not reflect their actual knowledge of the content, particularly when the assessment requires a lot of reading (Abedi, Leon, & Mirocha, 2003; Noble, Rosebery, Suarez, Warren, & O'Connor, 2014).

The ACT test measures college readiness in English, reading, math, and science (there is also an optional essay-format writing test, which was not considered in these analyses). The English and reading tests are completely text-based, whereas the math and science tests contain text as well as equations, figures, tables, and other illustrations. The ACT Composite score is the average of the four subject test scores, and the average Composite score of ACT-tested high school graduates in 2019 was 20.7 (ACT, 2019b). ACT has also developed College Readiness Benchmarks (ACT, 2019b), which are the scores at which a student has a 50% chance of earning a B or higher in a first-year, credit-bearing college course. The Benchmarks are 18 in English, 22 in math, 22 in reading, and 23 in science. These scores can be referred to when interpreting the results of this study with respect to the academic readiness of students in the study samples.

The testing supports provided to ELs on the ACT were reviewed by internal content experts as well as external experts in measurement, higher education, ELs, state and federal policy, and civil rights and were determined to be unlikely to alter the construct of interest and were generally supported in the literature as being effective (Abedi, Courtney & Leon, 2003; Abedi & Ewers, 2013; Abedi, Hofstetter, Baker, & Lord, 2001; Acosta, Rivera, Willner, 2008; Cawthon, Ho, Patel, Potvin, & Trundt, 2009; Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Kieffer, Rivera & Francis, 2012; Lovett, 2011; Pennock-Roman & Rivera, 2011).

Students can request supports when they register to take the ACT. Once the request is made, students work with school officials to provide relevant documentation to establish their EL status and ensure that they meet the requirements for use of the supports. ACT then makes a determination of whether the supports are approved for use when the students take the test.[1] In a National testing context, students or their parents are likely to be the parties making the decision to request the supports, whereas in a State and District in-school testing context, students may be receiving guidance from teachers or other staff about whether to request the supports.

The primary purpose of the ACT is to estimate the extent to which students are prepared for college-level coursework. Its primary uses are in college admissions, college course placement, scholarship eligibility, statewide assessment and accountability, and providing students with information about their relative strengths and areas to focus on for improvement. As a college readiness assessment, the ACT is designed to predict performance in college, where course content will be taught primarily in English. From this perspective, providing ELs with a test experience that allows them to demonstrate their knowledge and abilities in English while mitigating the effect of limited English proficiency on their scores in other subjects is appropriate in that scores are likely to reflect their level of college readiness and whether they might need additional supports once in college. While the test content of the ACT is not translated, test instructions are provided in 18 common languages (ACT, 2019a), and supports for ELs may include use of an approved word-to-word translation glossary containing no definitions.

This study provides evidence of the validity of ACT scores for ELs testing with supports. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." The Standards describe five types of validity evidence, including evidence based on test content, response processes, internal structure, relationships with other variables, and consequences of testing. The *ACT Technical Manual* (ACT, 2020) summarizes many years of research compiling validity evidence both overall and for several student subgroups, but given the recency of the policy change allowing ELs to test with supports, sufficient data has not been available to study the impact of supports for this population of students until this point. This study provides validity evidence based on internal structure, including classification consistency, differential item functioning (DIF), reliability, standard errors of measurement (SEM), and conditional standard errors of measurement (CSEM), as well as validity evidence based on relationships with other variables—in this case, self-reported high school grades.

Fairness of assessment requires that psychometric properties of scores be consistent across different subgroups of the test population. That is, measurement precision and consistency along with classification accuracy and consistency should be similar, and items should also function similarly across different subgroups. However, when score distributions of the subgroups differ substantially from each other, interpretations of the psychometric indices may be impacted.

Classification accuracy refers to the extent to which examinees would be classified into the same category based on observed test scores (i.e., the score an examinee would earn from a single test administration) and true scores (the unobserved theoretical score an examinee would earn if scores were averaged across an infinite number of test administrations). Classification consistency procedures estimate the

degree to which a student would be consistently classified into the same achievement level across multiple administrations of an assessment (Livingston & Lewis, 1995). Higher consistency is an indicator that the assessment has higher reliability for classification decisions. With strong assumptions about the distributions of true scores and measurement errors, these classification indices can be estimated using test scores from a single form.

DIF analyses provide estimates of the extent to which students in a focal group (in this case, ELs testing with supports or ELs testing without supports) have a different probability of answering a given test item correctly as compared to students in a reference group (in this case, non-Els testing without supports), given that the two groups have the same achievement level. A large amount of DIF favoring the reference group provides evidence that item bias may be present, putting the focal group at a disadvantage. It is standard practice for test developers to assess for DIF and revise or remove problematic items (ACT, 2020).

Reliability estimates the extent to which a student would have consistent scores across administrations of an assessment. CSEM provides estimates of score unreliability across the score scale. If CSEM differs for different groups, it is evidence that the assessment may be less precise for some groups and may be cause for concern.

High school grades are self-reported by students when they register to take the ACT. While there may be some limitations to using self-reported high school grades, previous research has found them to be highly accurate (Sanchez & Buddin, 2015). Students are also motivated to provide accurate information, as their grades may be verified by college personnel through official high school transcripts. Positive correlations are typically found between ACT scores and high school grades, as both are measures of academic performance.

The current study is one in a series of studies that will address the impact of providing testing supports to ELs taking the ACT. Additional research is needed to continue to compile other sources of validity evidence pertaining specifically to ELs testing with supports, and ACT has developed a research agenda to address the following validity claims.

Validity Claims

1. Providing supports to ELs who need the supports removes construct-irrelevant variance.

2. Providing supports to ELs who need the supports does not introduce item bias.

3. Providing supports to ELs who need the supports does not give them an unfair advantage.

Research Questions

1. How do ACT scores and score gains of ELs who initially tested without supports and retested with supports compare to the ACT scores and score gains of ELs and non-ELs who tested without supports?

2. How does the relationship between high school grades (HSGPA) and ACT scores for the three test groups compare across the two test events?

3. How do the scores of ELs who tested with or without supports compare to the scores of non-ELs who tested without supports with respect to classification consistency, DIF, reliability, and CSEM?

Methods

Samples

The gain score analysis sample (Research Questions 1 and 2) included students in the US who took the ACT twice within a 12-month window between September 2016 and July 2019. Students were excluded from the study sample if they did not test in the US (2%) or if state or district contract or law prohibited the use of their data for research purposes (15%). ELs who tested with supports were identified based on data from ACT's accommodations system. ELs who tested without supports and non-ELs were identified based on responses to a question presented when students registered to take the ACT: "Do you receive English language (EL) services at school now?" Students who responded "Yes" were classified as EL, and students who responded "No" were classified as non-EL. Students who responded "I prefer not to respond" were excluded from the analysis (18%). The resulting sample contained 2,279 ELs who tested without supports and retested with supports (0.2% of the study sample), 108,777 ELs who tested twice without supports (8% of the study sample), and 1,261,861 non-ELs who tested twice without supports (92% of the study sample).

A separate sample was used for the psychometric analyses (Research Question 3) to ensure that sufficient numbers of students were available to conduct the analyses. These analyses require large numbers of students taking a single test form of the ACT, and the sample of students who retested with supports was too small on any given test form to support these analyses; therefore, this sample was not restricted to students who tested more than once. Test forms were selected that included at least 1,000 students from each test group. The resulting sample came from six different administrations of the ACT in 2017–2019 and included 8,720 ELs who tested with

supports, 157,705 ELs who tested without supports, and 1,857,862 non-ELs who tested without supports.

Analyses

Performance and Score Gains

To determine whether ELs are benefitting from the supports, we examined ACT performance and score gains. Score gains were investigated both descriptively and using regressions models to control for number of months between tests and demographic covariates (income, parent education, and race/ethnicity). Three groups were compared: ELs who first tested without supports and then retested with supports, ELs who tested twice without any supports, and non-ELs who tested twice without any supports. In the regression models, the reference groups for the covariates were defined as follows: for income, the reference group was parent income above $36,000; for parent education, the reference group was non-first-generation college student (i.e., one parent or guardian had attended at least some college); and for race/ethnicity, the reference group was White students. Missing data was accounted for in the regression models by including dummy variables for the covariates that had missing data.

Disparities

A Disparity Index (DI) was also calculated to assess the extent to which performance gaps may have been reduced for ELs who retested with supports as compared to ELs who retested without supports. The DI can be interpreted as the percent difference in scores between two groups. It is calculated by subtracting the mean score of the reference group (non-EL) from the mean score of the focal group (EL), dividing the difference by the mean score of the focal group (EL), and multiplying by 100 (Abedi, 2002, 2009). A positive value indicates a performance gap favoring the focal group, and a negative value indicates a performance gap favoring the reference group.

Relationship with High School Grades

For those students who reported their high school courses taken and high school grades earned in those courses when they registered to take the ACT (68% of ELs who retested with supports, 84% of ELs who tested twice without supports, and 84% of non-ELs), correlation analyses were used to compare the relationship between high school grades for each test event by testing group. If the testing supports allow ELs to more accurately demonstrate their academic abilities, and assuming that high school grades are an accurate measure of academic ability, the correlations should be higher for the test in which they tested with supports.

Score Comparability

Score Comparability. Score comparability was evaluated with respect to classification consistency, DIF, reliability, SEM, and CSEM. Three groups of interest were compared: ELs who tested with supports, ELs who tested without supports, and non-ELs who tested without supports.

Classification consistency values were computed based on the ACT College Readiness Benchmarks. Classification accuracy and classification consistency rates were calculated based on the method described by Livingston and Lewis (1995), which assumes that the true scores are distributed following a four-parameter beta distribution and that conditional errors are distributed following a binomial distribution. Kappa statistics were also calculated for classification accuracy and consistency. These statistics take into account the chance agreement rate.

DIF analyses were performed using Mantel-Haenszel (MH) procedures. ELs who tested with supports and ELs who tested without supports were the focal groups, and non-ELs who tested without supports were the reference group.

Mantel-Haenszel Chi-square (MH-CHISQ) p-values and Mantel-Haenszel effect sizes (MH-D) were used to classify the items into different categories (A: negligible DIF, B: moderate DIF, or C: large DIF) following the criteria listed in Table 14. Plus and minus signs were used to indicate whether the item favors the focal group (+) or the reference group (-).

Standard errors of measurement (SEM) and reliability are commonly used indices for measurement precision and consistency. SEM is in the unit of the reported score scale and can be used to construct a confidence interval for the observed scores. If SEM varies across different levels of the score scale, conditional SEM (CSEM) can be used as a more accurate indication of individual level score precision. Reliability, with a range of 0 to 1, is an index of score consistency with repeated measures.

A strong true score model (Lord, 1965) was used to estimate the CSEM, SEM, and reliability of the ACT scale scores for each test form and subject area. True number correct scores were assumed to have a four-parameter beta distribution and measurement errors conditional on each true score were assumed to have a compound binomial distribution. Scale score CSEM, SEM, and reliability were estimated following procedures described in Kolen, Hanson, and Brennan (1992).

Results

Table 1 contains demographic characteristics of the retest sample. It should first be noted that ELs who retested with supports were more likely to have not provided demographic information as compared to the other two study groups. It is unknown

whether this is due to a language barrier when registering to take the ACT or some other factor. ELs who retested with supports contained larger proportions of Hispanic/Latino and Asian students, ELs who tested twice without supports contained larger proportions of Black/African American students, and non-ELs contained larger proportions of White students. ELs were also more likely to report low income, defined as parent income of less than $36,000 per year. ELs who retested with supports were the most likely to report low income, followed by ELs who tested without supports, and non-ELs were the least likely to report low income. ELs were also more likely to be first-generation college students, defined as neither parent having had any college. ELs who retested with supports were the most likely to be first-generation college students, followed by ELs who tested without supports, and non-ELs were the least likely to be first-generation college students.

Table 1. Demographic Characteristics (Percentages) and Size of Score Gain Study Sample

Demographic Characteristic	EL Retested with Supports	EL Tested Twice without Supports	Non-EL without Supports
Race/Ethnicity			
Black/African American	13	23	10
American Indian/Alaska Native	0	1	1
White	6	49	61
Hispanic/Latino	48	14	11
Asian	23	5	6
Native Hawaiian/Pacific Islander	0	0	0
Two or More Races	0	3	4
No Response/Missing	9	5	6
Income			
Low Income < $36,000	34	24	12
Not Low Income	14	46	58
No Response/Missing	52	30	30
Parent Education			
First Generation College Student	31	19	10
Not First Generation	30	64	76
No Response/Missing	39	17	15
Sample Size	2,279	108,777	1,261,861

Research Question 1: How do ACT scores and score gains of ELs who initially tested without supports and retested with supports compare to the ACT scores and score gains of ELs and non-ELs who tested without supports?

Initial Scores

To make a preliminary determination of the fairness and appropriateness of providing EL supports, we compared each group's ACT English scores on their first test which was administered without supports (Table 2). ELs who retested with supports demonstrated substantially lower English performance (13.0) than ELs who received no supports (17.9) and non-ELs (22.0). While English ACT scores were not the basis for selecting ELs to receive supports on retest, their performance suggests that ELs who tested twice without supports had higher levels of English proficiency than ELs who retested with supports. ELs who tested twice without supports scored near the ACT College Readiness Benchmark of 18 in English on their first test attempt (17.9) and above the Benchmark on their second test attempt (19.1), suggesting that on average, their English proficiency is sufficient to succeed in a first-year college English composition course.

Performance and Score Gains

Table 2 contains mean scores and score gains for the three test groups. Comparing average ACT Composite scores on the first test, ELs who retested with supports had the lowest performance (15.1), followed by ELs who tested twice without supports (18.8); non-ELs had the highest performance (22.4).

Table 2. Mean ACT Scores and Score Gains by Test Group and Disparity Indices (DI) Relative to Non-ELs

	First Test			Second Test			Gain	
EL Retested with Supports	Mean	SD	DI	Mean	SD	DI	Mean	SD
English	13.0	3.9	-70.0	14.7	5.1	-60.1	1.7	3.3
Math	16.9	4.5	-29.7	17.8	5.4	-27.1	0.9	2.4
Reading	14.3	3.7	-60.8	16.6	5.2	-43.6	2.4	4.4
Science	15.9	4.1	-40.0	16.9	5.3	-35.5	1.0	4.0
Composite	15.1	3.4	-48.1	16.6	4.6	-40.4	1.5	2.4
EL Tested Twice without Supports	First Test			Second Test			Gain	
	Mean	SD	DI	Mean	SD	DI	Mean	SD
English	17.9	5.6	-23.3	19.1	6.1	-22.8	1.3	3.1
Math	18.8	4.5	-16.1	19.4	4.9	-16.6	0.5	2.4
Reading	18.8	5.4	-21.8	19.7	5.8	-21.4	0.8	3.8
Science	19.1	4.5	-16.1	19.6	4.8	-16.5	0.5	3.4
Composite	18.8	4.5	-19.1	19.6	4.9	-19.2	0.8	2.0
Non-EL without Supports	First Test			Second Test			Gain	
	Mean	SD		Mean	SD		Mean	SD
English	22.0	6.2		23.5	6.5		1.5	3.2
Math	21.9	5.1		22.6	5.4		0.7	2.5
Reading	22.9	6.2		23.9	6.4		0.9	3.9
Science	22.2	4.9		22.9	5.2		0.7	3.4
Composite	22.4	5.0		23.3	5.3		1.0	2.0

Note: The DI is calculated by subtracting the mean score of the reference group (non-EL) from the mean score for the focal group (EL), dividing the difference by the mean score for the focal group (EL), and multiplying by 100, and can be interpreted as the percent difference in scores between the focal and reference groups (Abedi, 2002, 2009).

ELs who retested with supports tended to score lowest in English (13.0) and highest in math (16.9) on their first test attempt, whereas non-ELs tended to score highest in reading (22.9) and lowest in math (21.9). Comparing the first test scores of ELs who retested with supports to those of non-ELs, the largest gap was in English (9.1 points, or 1.5 standard deviations) and the smallest gap was in math (5.0 points, or 1.0 standard deviations).

ELs who retested with supports showed the largest Composite score gains, gaining 1.5 points compared to 0.8 for ELs who tested twice without supports and 1.0 for non-ELs. Score gains of ELs who retested with supports were the highest in reading at 2.4 points; score gains in other subject areas were slightly higher but similar to those of non-ELs. The larger gains of ELs who retested with supports suggests that the supports do, indeed, provide a benefit to EL students.

Disparities

Table 2 also contains the Disparity Index (DI) values for the two EL groups compared to non-ELs. Comparing the two EL groups, ELs who retested with supports had larger disparities on both test events and across all subject areas than ELs who tested twice without supports. Comparing the first test DI value to the second test DI value, the disparities were similar (within 0.5%) across test events for ELs who tested twice without supports. However, the disparities were reduced upon retest for ELs who retested with supports, particularly in English (disparity reduced by 9.9%) and reading (disparity reduced by 17.1%). For both EL groups, the disparities were larger in English and reading for both test events.

Regression Analyses

Regression Analyses. Because ELs differ from non-ELs with respect to demographic characteristics and to account for the confounding effects of any learning gains that may occur as the number of months between test events increases, regression analyses were conducted that included relevant covariates to understand the impact of EL supports on ACT score gains controlling for these factors. Results are presented in Tables 3–7. All else being equal, the results found that EL students testing with supports had ACT Composite score gains that were 0.4 points higher than those for non-EL students, and ELs who tested without supports had Composite score gains that were 0.2 points lower than those for non-EL students. These estimated gains were similar to the findings from the descriptive analyses (Table 2), where Composite score gains for ELs who retested with supports were 0.5 points higher than gains of non-ELs (1.5–1.0), and Composite score gains for ELs who tested twice without supports were 0.2 points lower than gains of non-ELs (0.8–1.0).

Table 3. Regression Predicting Gains in ACT Composite Score

Variable	Beta	SE	t Value	Pr > |t|
Intercept	1.466	0.010	146.15	<.0001
ACT English (First Test)	0.032	0.000	65.00	<.0001
ACT Math (First Test)	0.069	0.001	123.28	<.0001
ACT Reading (First Test)	-0.059	0.000	-126.35	<.0001
ACT Science (First Test)	-0.075	0.001	-122.33	<.0001
Months between Tests	0.081	0.001	154.58	<.0001
EL Retested with Supports	0.400	0.041	9.82	<.0001
EL Tested Twice without Supports	-0.233	0.006	-37.16	<.0001
Low Income (Less than $36,000)	-0.222	0.006	-40.10	<.0001
Income Missing	0.053	0.005	11.72	<.0001
First Generation College Student	-0.288	0.006	-48.43	<.0001
Parent Education Missing	-0.144	0.006	-24.73	<.0001
Black	-0.516	0.006	-89.41	<.0001
American Indian	-0.331	0.020	-16.40	<.0001
Hispanic	-0.273	0.006	-49.45	<.0001
Asian	0.036	0.007	4.94	<.0001
Pacific Islander	-0.214	0.040	-5.41	<.0001
Multiple Races/Ethnicities	-0.126	0.008	-14.89	<.0001
Missing Race/Ethnicity	-0.124	0.007	-17.07	<.0001

For the subject-specific regression analyses (Tables 4–7), as we would expect, examinees' initial scores in that subject area had a negative relationship with their gain scores in that subject area, and this holds across subject areas. That is, students with higher initial scores showed smaller gains upon retest, and students with lower initial scores showed larger gains upon retest (Camara & Allen, 2017). Interestingly, after accounting for months between test events and demographic characteristics, the estimated score gains for ELs who retested with supports were not statistically significantly different from those of non-ELs in English and science. After controlling for relevant covariates, the estimated gains of ELs who retested with supports were 0.7 points higher than the gains of non-ELs in math (compared to 0.2 points before controlling for covariates) and 0.9 points higher in reading (compared to 1.4 points before controlling for covariates).

Table 4. Regression Predicting Gains in ACT English Score

Variable	Beta	SE	t Value	Pr > |t|
Intercept	0.246	0.015	16.08	<.0001
ACT English (First Test)	-0.353	0.001	-465.86	<.0001
ACT Math (First Test)	0.166	0.001	194.93	<.0001
ACT Reading (First Test)	0.150	0.001	210.32	<.0001
ACT Science (First Test)	0.069	0.001	73.57	<.0001
Months between Tests	0.110	0.001	136.89	<.0001
EL Retested with Supports	-0.075	0.062	-1.20	0.2305
EL Tested Twice without Supports	-0.256	0.010	-26.82	<.0001
Low Income (Less than $36,000)	-0.306	0.008	-36.09	<.0001
Income Missing	0.135	0.007	19.43	<.0001
First Generation College Student	-0.485	0.009	-53.37	<.0001
Parent Education Missing	-0.263	0.009	-29.58	<.0001
Black	-0.562	0.009	-63.78	<.0001
American Indian	-0.622	0.031	-20.16	<.0001
Hispanic	-0.402	0.008	-47.57	<.0001
Asian	-0.001	0.011	-0.13	0.8981
Pacific Islander	-0.297	0.060	-4.92	<.0001
Multiple Races/Ethnicities	-0.199	0.013	-15.41	<.0001
Missing Race/Ethnicity	-0.157	0.011	-14.21	<.0001

Table 5. Regression Predicting Gains in ACT Math Score

Variable	Beta	SE	t Value	Pr > |t|
Intercept	1.142	0.012	93.56	<.0001
ACT English (First Test)	0.085	0.001	140.05	<.0001
ACT Math (First Test)	-0.282	0.001	-414.54	<.0001
ACT Reading (First Test)	0.006	0.001	9.86	<.0001
ACT Science (First Test)	0.159	0.001	212.70	<.0001
Months between Tests	0.055	0.001	85.33	<.0001
EL Retested with Supports	0.701	0.050	14.13	<.0001
EL Tested Twice without Supports	-0.097	0.008	-12.72	<.0001
Low Income (Less than $36,000)	-0.245	0.007	-36.26	<.0001
Income Missing	0.084	0.006	15.13	<.0001
First Generation College Student	-0.241	0.007	-33.25	<.0001
Parent Education Missing	-0.144	0.007	-20.34	<.0001
Black	-0.459	0.007	-65.33	<.0001
American Indian	-0.344	0.025	-13.98	<.0001
Hispanic	-0.207	0.007	-30.80	<.0001
Asian	0.491	0.009	55.34	<.0001
Pacific Islander	-0.056	0.048	-1.16	0.2466
Multiple Races/Ethnicities	-0.142	0.010	-13.82	<.0001
Missing Race/Ethnicity	-0.058	0.009	-6.56	<.0001

Table 6. Regression Predicting Gains in ACT Reading Score

Variable	Beta	SE	t Value	Pr > |t|
Intercept	1.439	0.018	80.88	<.0001
ACT English (First Test)	0.288	0.001	327.09	<.0001
ACT Math (First Test)	0.057	0.001	57.11	<.0001
ACT Reading (First Test)	-0.527	0.001	-636.43	<.0001
ACT Science (First Test)	0.161	0.001	147.67	<.0001
Months between Tests	0.098	0.001	104.87	<.0001
EL Retested with Supports	0.913	0.072	12.63	<.0001
EL Tested Twice without Supports	-0.384	0.011	-34.58	<.0001
Low Income (Less than $36,000)	-0.128	0.010	-12.99	<.0001
Income Missing	0.011	0.008	1.31	0.1911
First Generation College Student	-0.212	0.011	-20.07	<.0001
Parent Education Missing	-0.091	0.010	-8.77	<.0001
Black	-0.409	0.010	-40.01	<.0001
American Indian	-0.151	0.036	-4.22	<.0001
Hispanic	-0.097	0.010	-9.89	<.0001
Asian	-0.266	0.013	-20.56	<.0001
Pacific Islander	-0.253	0.070	-3.61	0.0003
Multiple Races/Ethnicities	-0.004	0.015	-0.25	0.7994
Missing Race/Ethnicity	-0.101	0.013	-7.82	<.0001

Table 7. Regression Predicting Gains in ACT Science Score

Variable	Beta	SE	t Value	Pr > |t|
Intercept	3.068	0.015	209.02	<.0001
ACT English (First Test)	0.109	0.001	150.53	<.0001
ACT Math (First Test)	0.334	0.001	409.38	<.0001
ACT Reading (First Test)	0.136	0.001	198.75	<.0001
ACT Science (First Test)	-0.692	0.001	-769.71	<.0001
Months between Tests	0.065	0.001	84.19	<.0001
EL Retested with Supports	0.058	0.060	0.98	0.3282
EL Tested Twice without Supports	-0.192	0.009	-20.92	<.0001
Low Income (Less than $36,000)	-0.212	0.008	-26.07	<.0001
Income Missing	-0.011	0.007	-1.73	0.0843
First Generation College Student	-0.221	0.009	-25.39	<.0001
Parent Education Missing	-0.089	0.009	-10.46	<.0001
Black	-0.638	0.008	-75.61	<.0001
American Indian	-0.232	0.030	-7.86	<.0001
Hispanic	-0.387	0.008	-47.88	<.0001
Asian	-0.079	0.011	-7.39	<.0001
Pacific Islander	-0.242	0.058	-4.18	<.0001
Multiple Races/Ethnicities	-0.152	0.012	-12.34	<.0001
Missing Race/Ethnicity	-0.177	0.011	-16.63	<.0001

Research Question 2: How does the relationship between high school grades (HSGPA) and ACT scores for the three test groups compare across the two test events?

Table 8 contains the average self-reported high school grades of students by test group. Both overall and for each subject area, ELs who retested with supports had the lowest grades, followed by ELs who tested without supports, and non-ELs had the highest grades, mirroring the patten observed by ACT test scores. Students in all three groups had lower grades in math and science than in English and social studies, whereas ELs who retested with supports had higher ACT scores in math and science than in English and reading (Table 2).

Table 8. Average High School Grades by Test Group

	EL Retested with Supports		EL Tested Twice without Supports		Non-EL without Supports	
Subject Area	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
English	3.23	0.71	3.41	0.63	3.61	0.53
Math	3.22	0.77	3.28	0.71	3.48	0.61
Social Sciences	3.27	0.71	3.48	0.62	3.65	0.50
Science	3.19	0.72	3.35	0.66	3.54	0.56
Overall	3.24	0.60	3.37	0.55	3.57	0.46
Sample Size	1,537		91,256		1,061,696	

Table 9 contains correlations between examinees' ACT scores and high school grades for each test attempt by subject area and test group. For all three test groups, the correlations between HSGPA and ACT scores were slightly higher on their second test attempt. The increase in correlations for ELs who tested without supports and non-ELs tended to be 0.01 to 0.02, whereas the increase in correlations for ELs who retested with supports were higher in reading (0.05) and science (0.06), suggesting that the supports were indeed allowing them to better demonstrate their true achievement level in these subjects.

Table 9. Correlations between ACT Scores and High School Grades, by Subject Area and Test Group

	ACT	English	Math	Reading	Science	Composite
	Grades	**English**	**Math**	**Social Studies**	**Science**	**Overall**
EL Retested with Supports	Test 1	0.24	0.42	0.24	0.27	0.39
	Test 2	0.26	0.44	0.29	0.34	0.43
EL Tested Twice without Supports	Test 1	0.42	0.46	0.36	0.40	0.52
	Test 2	0.44	0.48	0.37	0.41	0.53
Non-EL without Supports	Test 1	0.41	0.48	0.35	0.41	0.52
	Test 2	0.43	0.50	0.36	0.42	0.53

Note: All correlations were significant at $p < 0.0001$.

Research Question 3: How do the scores of ELs who tested with or without supports compare to the scores of non-ELs who tested without supports with respect to classification consistency, DIF, reliability, and CSEM?

Score comparability was evaluated using a different sample than that used for the score gain analyses, as described earlier. Three test groups were compared: ELs who tested with supports, ELs who tested without supports, and non-ELs who tested without supports. Descriptive statistics of the data sample for the three groups of interest are provided in Table 10, and plots of the relative frequency distributions of ACT Composite scores are provided in Figure 1. The large differences in score distributions among the groups should be kept in mind when interpreting the results of the following analyses. Six different test forms were used to create the data sample. Table 11 provides the sample sizes for each form. Subsequent analyses were conducted by test form and subject area.

Table 10. Sample Size and Mean ACT Scores for Psychometric Analysis Sample by Test Group

	EL with Supports		EL, No Supports		Non-EL, No Supports	
Subject Area	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
English	12.8	4.0	18.0	6.2	22.5	6.7
Math	16.0	4.1	18.7	4.8	21.9	5.4
Reading	14.9	4.4	19.0	5.8	23.2	6.5
Science	15.4	4.3	19.0	4.8	22.2	5.3
Composite	14.9	3.5	18.8	4.8	22.6	5.4
Sample Size	8,720		157,705		1,857,862	

Table 11. Sample Size by Test Form

Form	EL with Supports	EL, No Supports	Non-EL, No Supports	Total
1	1,276	28,735	315,155	345,166
2	1,920	3,091	12,150	17,161
3	2,207	41,536	490,354	534,097
4	1,056	39,885	534,786	575,727
5	1,066	25,170	288,752	314,988
6	1,195	19,288	216,665	237,148
Total	8,720	157,705	1,857,862	2,024,287

Figure 1. Distribution of ACT Composite Scores by Test Group

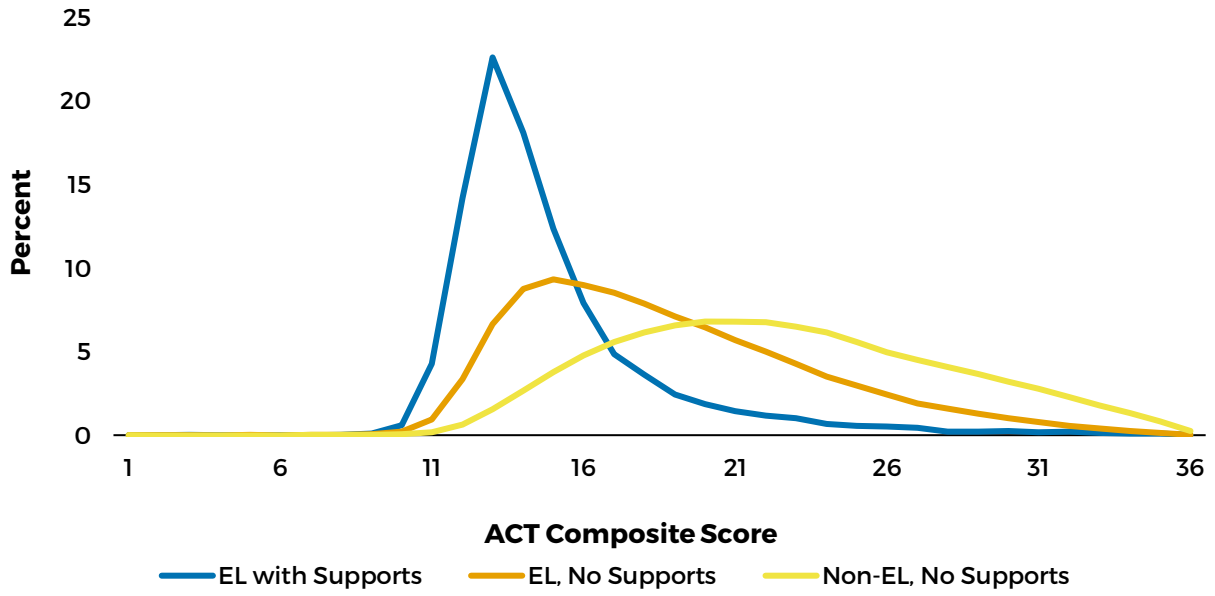


Table 12 contains classification accuracy consistency results by test form and test group. Across all subject areas, test forms, and test groups, classification accuracy and consistency rates were high (0.82–0.98). Classification consistency rates were the highest for ELs who tested with supports (0.94–0.98), followed by ELs who tested without supports (0.86–0.96); non-ELs who tested without supports had slightly lower classification consistency rates (0.82–0.92) than the other two groups. The classification accuracy rates showed a similar pattern. Note that the percentage of agreement in Table 12 is influenced by the relative position of the cut scores in the score distribution. When the cut scores are toward the end of the distribution where frequencies are low, percentage of consistent classifications is expected to be higher than when the cut scores are at the middle of the distribution where frequencies are high. This is why the EL group with supports had the highest agreement rates, as ELs, particularly those who tested with supports, tended to score far below the cut scores.

Table 12. Classification Accuracy and Consistency by Test Group

Subject (Cut Score)	Form	EL with Supports	EL, No Supports	Non-EL, No Supports	EL with Supports	EL, No Supports	Non-EL, No Supports
English (18)	1	0.97	0.89	0.87	0.96	0.89	0.88
	2	0.96	0.92	0.89	0.96	0.92	0.89
	3	0.95	0.90	0.89	0.95	0.90	0.89
	4	0.95	0.88	0.87	0.95	0.89	0.88
	5	0.95	0.89	0.88	0.94	0.89	0.88
	6	0.96	0.90	0.88	0.96	0.90	0.89
Math (22)	1	0.97	0.93	0.88	0.97	0.93	0.88
	2	0.98	0.97	0.92	0.98	0.96	0.92
	3	0.98	0.94	0.88	0.98	0.94	0.88
	4	0.96	0.91	0.87	0.97	0.91	0.87
	5	0.96	0.94	0.89	0.97	0.93	0.88
	6	0.97	0.95	0.90	0.98	0.94	0.90
Reading (22)	1	0.95	0.88	0.85	0.95	0.88	0.85
	2	0.97	0.93	0.90	0.98	0.93	0.89
	3	0.94	0.90	0.85	0.95	0.89	0.85
	4	0.96	0.88	0.85	0.96	0.88	0.86
	5	0.96	0.89	0.84	0.96	0.89	0.84
	6	0.95	0.88	0.83	0.95	0.88	0.83
Science (23)	1	0.96	0.87	0.83	0.96	0.86	0.82
	2	0.97	0.92	0.88	0.97	0.92	0.87
	3	0.95	0.89	0.83	0.96	0.89	0.83
	4	0.95	0.87	0.84	0.95	0.86	0.84
	5	0.94	0.88	0.83	0.94	0.87	0.82
	6	0.97	0.91	0.86	0.97	0.90	0.86

Table 13 presents Kappa statistics, which are the classification accuracy and consistency results after accounting for agreement by chance. These values tended to be similar across the three test groups except for in math. The kappa statistics for ELs who tested with supports were slightly higher in math than those for the other two test groups.

Table 13. Kappa Statistics for Classification Accuracy and Consistency by Test Group

Subject (Cut Score)	Form	EL with Supports	EL, No Supports	Non-EL, No Supports	EL with Supports	EL, No Supports	Non-EL, No Supports
English (18)	1	0.73	0.77	0.72	0.74	0.77	0.71
	2	0.73	0.77	0.78	0.73	0.77	0.78
	3	0.81	0.79	0.73	0.80	0.80	0.72
	4	0.81	0.77	0.65	0.80	0.76	0.63
	5	0.78	0.77	0.72	0.79	0.77	0.71
	6	0.79	0.78	0.74	0.77	0.78	0.72
Math (22)	1	0.89	0.80	0.76	0.86	0.81	0.76
	2	0.88	0.77	0.77	0.83	0.78	0.78
	3	0.91	0.81	0.76	0.90	0.82	0.76
	4	0.87	0.79	0.73	0.85	0.79	0.73
	5	0.88	0.80	0.76	0.86	0.81	0.76
	6	0.91	0.82	0.79	0.84	0.83	0.79
Reading (22)	1	0.64	0.69	0.69	0.64	0.69	0.69
	2	0.70	0.69	0.75	0.66	0.69	0.76
	3	0.75	0.73	0.70	0.74	0.74	0.69
	4	0.73	0.73	0.71	0.70	0.73	0.70
	5	0.79	0.72	0.68	0.79	0.72	0.68
	6	0.71	0.69	0.67	0.72	0.70	0.66
Science (23)	1	0.68	0.60	0.64	0.70	0.61	0.64
	2	0.54	0.49	0.65	0.52	0.50	0.65
	3	0.74	0.65	0.66	0.69	0.65	0.66
	4	0.69	0.65	0.68	0.67	0.66	0.68
	5	0.72	0.60	0.63	0.72	0.62	0.64
	6	0.63	0.66	0.71	0.59	0.66	0.71

DIF analyses were conducted using a total of 1,290 items from six test forms. Criteria for negligible (A), moderate (B), and large (C) DIF are presented in Table 14. Comparing the ELs without supports (focal group) to the non-ELs (reference group), only one item (which was an English item) was identified as a B- DIF item, indicating that the items function the same between ELs who tested without supports and non-ELs. Between ELs who tested with supports and non-ELs, however, many items were classified as DIF items. Table 15 presents the number of items classified as A, B+, B-, C+, and C- across the six test forms overall and for each subject area. B+ and C+ indicate DIF items favoring the focal group, and B- and C- indicate DIF items favoring the reference group.

Table 14. Criteria for A, B, and C DIF Categories for the MH Procedure for MC Items

Category	Description	Criterion
A	Negligible DIF	Nonsignificant MH-CHISQ (P > 0.05) or |MH-D| < 1.0
B	Moderate DIF	Nonsignificant MH-CHISQ (P > 0.05) and 1.0 ≤ |MH-D| < 1.5
C	Large DIF	Nonsignificant MH-CHISQ (P > 0.05) and |MH-D| ≥ 1.5

Table 15. Summary of DIF Classifications between ELs with Supports (Focal) and Non-ELs (Reference)

Subject	A	B+	B-	C+	C-	Total	Flagged	% Flagged	%+	%-
English	318	37	33	33	29	450	132	29%	16%	14%
Math	300	20	21	4	15	360	60	17%	7%	10%
Reading	183	21	14	9	13	240	57	24%	13%	11%
Science	207	14	11	1	7	240	33	14%	6%	8%
Total	1,008	92	79	47	64	1,290	282	22%	11%	11%

Note: + indicates DIF favoring focal group and - indicates DIF favoring reference group.

Overall, 78% of items were not flagged for DIF, 13% were flagged for moderate DIF, and 9% were flagged for large DIF. The last two columns of Table 15 contain the percentages of items favoring the focal group ("%+") and the percentages of items favoring the reference group ("%-"). Note that the percentages of items favoring the reference group tended to be similar to those favoring the focal group, indicating that the impact of these DIF items on the total test scores should be minimal.

Since the EL with supports and non-EL groups had large differences in both score distributions and sample sizes, it was not known whether the flagged items were spurious DIF items due to these differences. Further investigations were conducted using matched samples from the non-EL group using sample sizes of 1, 5, and 10 times that of the ELs with supports group. The percentage of flagged items in these further analyses using the matched samples did not differ much from the original results.

Figures 2–5 contain the CSEM of scale scores for the three test groups by subject area and test form. For these figures, group 0 represents non-ELs, group 1 represents ELs with supports, and group 2 represents ELs without supports. The curves for the three test groups are virtually indistinguishable, overlapping across the entire scale score range for each subject area and each test form, which indicates that measurement precision was comparable across the three test groups at each true scale score.

Figure 2. Conditional Standard Errors of Measurement for ACT English by Test Group

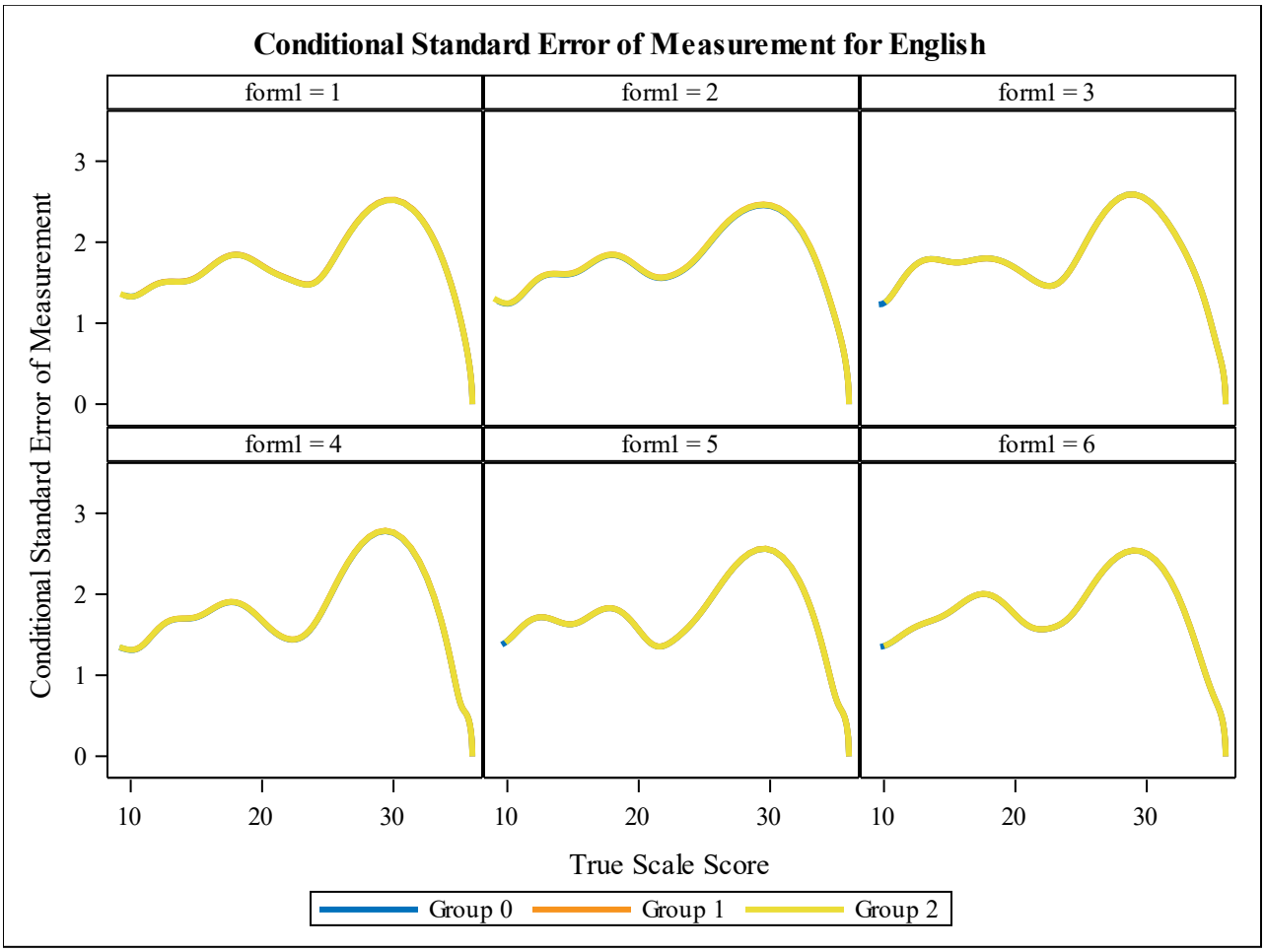


Figure 3. Conditional Standard Errors of Measurement for ACT Math by Test Group

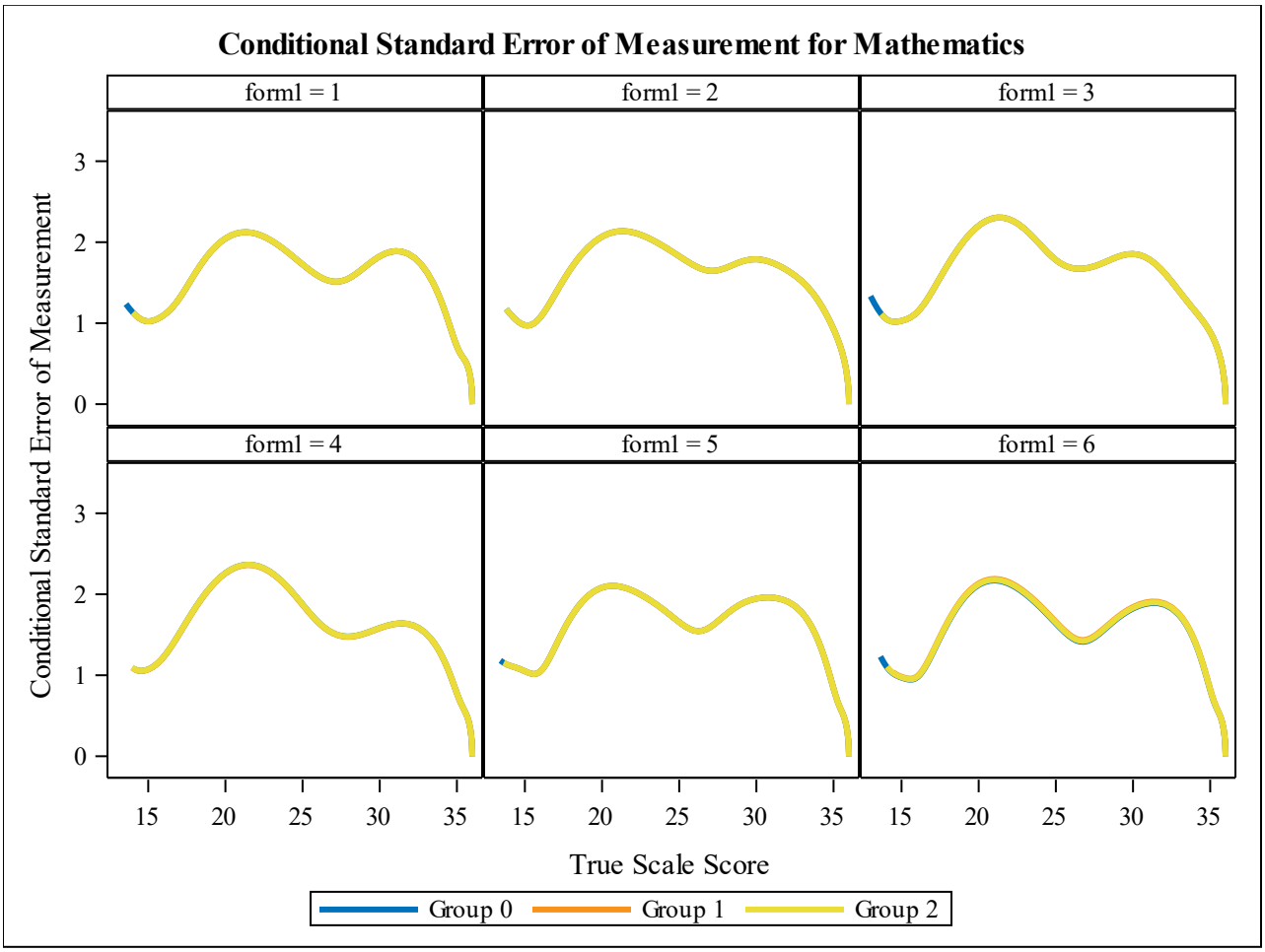


Figure 4. Conditional Standard Errors of Measurement for ACT Reading by Test Group

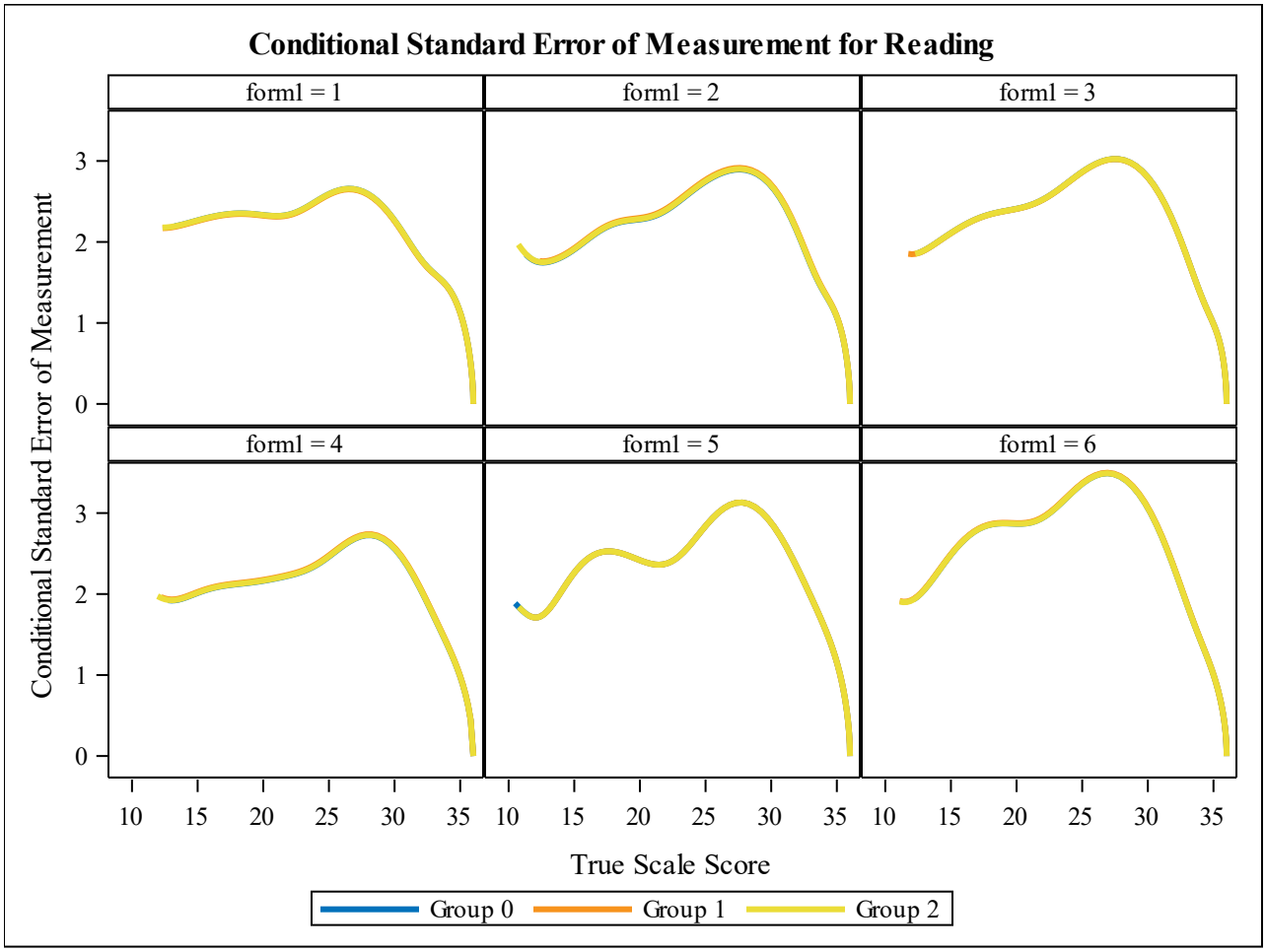
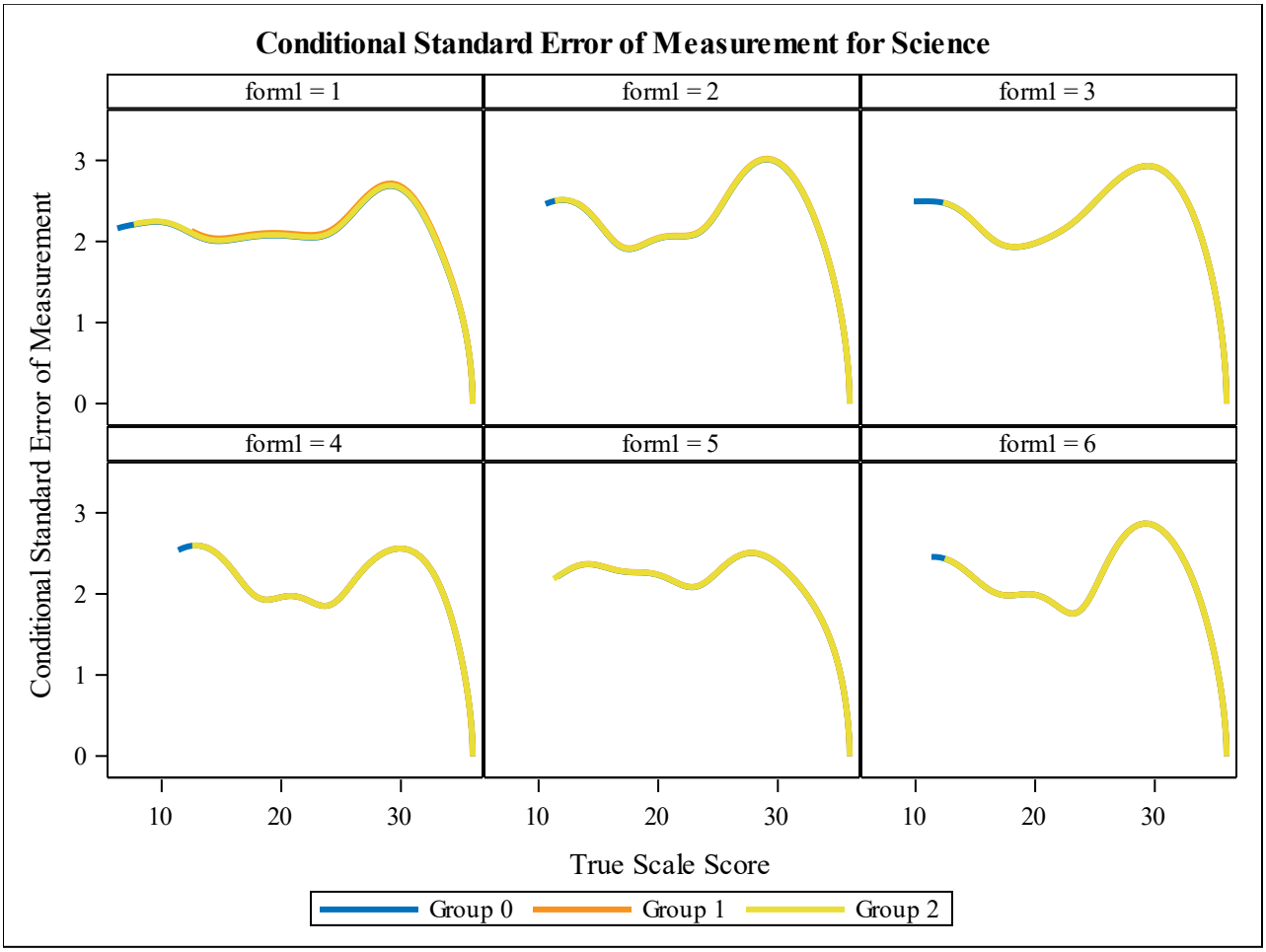


Figure 5. Conditional Standard Errors of Measurement for ACT Science by Test Group



Even though the CSEMs were identical across the three test groups, scale score reliability and SEM could differ due to differences in the true score distributions of the groups. Table 16 presents the scale score reliability and SEM for each subject area and test form. Except for math, the reliability estimates tended to be slightly lower for ELs with supports than for ELs without supports and non-ELs. Comparisons where the differences in reliability estimates were larger than 0.05 between the two EL groups are in bold. SEMs for ELs with supports, however, all tended to be lower than those of the other two groups except for in science. The reliability and SEM values were all similar between ELs without supports and non-ELs.

Table 16. Reliability and SEM of Scale Scores

		Reliability			SEM		
Subject	**Form**	**EL with Supports**	**EL, No Supports**	**Non-EL, No Supports**	**EL with Supports**	**EL, No Supports**	**Non-EL, No Supports**
English	1	**0.80**	0.91	0.92	1.46	1.65	1.76
	2	**0.80**	0.90	0.92	1.43	1.57	1.70
	3	0.88	0.92	0.93	1.53	1.71	1.82
	4	0.89	0.92	0.92	1.52	1.79	1.88
	5	0.88	0.91	0.92	1.62	1.69	1.76
	6	**0.83**	0.91	0.93	1.55	1.74	1.84
Math	1	0.91	0.89	0.90	1.19	1.50	1.66
	2	0.87	0.84	0.88	1.19	1.36	1.56
	3	0.92	0.89	0.90	1.22	1.55	1.76
	4	0.91	0.89	0.89	1.34	1.70	1.81
	5	0.92	0.89	0.90	1.31	1.50	1.68
	6	0.90	0.90	0.92	1.08	1.41	1.61
Reading	1	**0.71**	0.82	0.85	2.25	2.32	2.33
	2	**0.68**	0.81	0.86	1.88	2.06	2.24
	3	0.81	0.83	0.86	2.12	2.33	2.45
	4	**0.77**	0.86	0.87	2.02	2.18	2.23
	5	0.82	0.84	0.85	2.07	2.36	2.51
	6	0.75	0.79	0.83	2.40	2.69	2.81
Science	1	0.76	0.79	0.81	2.11	2.10	2.16
	2	**0.61**	0.72	0.80	2.31	2.18	2.17
	3	**0.73**	0.79	0.82	2.26	2.20	2.31
	4	**0.73**	0.81	0.84	2.38	2.13	2.09
	5	0.80	0.78	0.80	2.29	2.27	2.25
	6	**0.66**	0.80	0.86	2.25	2.13	2.14

Note: Reliability estimates differing by more than 0.05 between the two EL groups are in bold.

Discussion

This study investigated the performance of ELs taking the ACT using testing supports, including analyses of score gains and disparities, relationships between scores and high school grades, and score comparability. We found that Els overall tended to score substantially below non-ELs, and ELs who took the ACT with testing supports tended to have much lower scores than ELs who tested without supports. On average, ACT-tested students tended to gain about one score point upon retest, and ELs who first tested without supports and retested with supports tended to gain about 1.5 score points, with higher gains in English (1.7) and reading (2.4).

Performance gaps for ELs who retested with supports were reduced but not eliminated upon retesting. Given the data available for this study, it is unknown the extent to which ELs' performance was due to limited English proficiency or to actually having lower levels of academic achievement, possibly due to interruptions in education, less access to core academic coursework, or income or race-related inequities (Callahan & Shifrer, 2016; Johnson, 2019; Moore, in press; Sugarman, 2019). While ELs who retested with supports did show higher gains than the other two test groups, the gains were not unreasonable, being within one standard deviation, and ELs retesting with supports still scored substantially below non-ELs, suggesting that the supports were not providing an unfair advantage.

We also found that correlations between high school grades and ACT scores were slightly higher on the second test attempt across the three test groups, but the increase in correlations were higher in reading and science for ELs who retested with supports, suggesting that the supports did indeed remove some construct irrelevant variance from their scores, providing initial convergent validity evidence.

Psychometric properties of test scores were also examined using a separate sample of ELs who tested with supports, ELs who tested without supports, and non-ELs who tested without supports. Results of these analyses indicated that conditional measurement precision (i.e., CSEM) for ELs who tested with supports was equivalent across the three test groups. Reliability and SEM estimates as well as classification accuracy and consistency indices, however, showed some differences among the groups. Lower score reliability was found for ELs who tested with supports, which is consistent with other studies examining reliability across different assessments, subject areas, and grade levels (Moore, Li, & Lu, 2020). Paradoxically, ELs who tested with supports tended to have lower reliability but at the same time lower SEM. These differences and seemingly paradoxical observations were likely due to the differences in true score distributions between the three groups.

DIF was investigated using the Mantel-Haenszel procedure. Though only a single item was identified as exhibiting DIF between the ELs who tested without supports and non-ELs, an average of 22% of items across all subject areas and all forms were flagged as exhibiting DIF. Among these items, however, about half favored the focal group and the other half favored the reference group, meaning that the impact of these items on the total test scores may be minimal. Additional analyses were conducted using matched samples, and the results suggested that the majority of the items identified as exhibiting DIF were not explained by differences in score distributions or sample sizes. Further investigations should be conducted to better understand the source and the impact of DIF.

The results of this study provide preliminary evidence that allowing ELs to use testing supports when taking the ACT is benefiting this population of students without

conferring an unfair advantage. Because one of the primary uses of the ACT is to assess college readiness, one of the most important pieces of validity evidence will ultimately be the impact of these supports on predicting college performance. To that end, ACT is currently recruiting colleges to participate in a predictive validity study in which colleges will provide students' first-year college grades with the ultimate goal of evaluating whether the ACT scores of ELs testing with supports are accurate predictors of college performance. While the Covid-19 pandemic has postponed data collection for this study, we are hopeful that this research will resume in 2021.

Study Significance

This study is one of a series of studies examining the effects of providing test supports to ELs on their ACT scores. It is important that we investigate and document the impact of offering supports to ELs on the ACT in terms of reliability, validity, and fairness of scores. Understanding the impact of providing test supports is essential for making appropriate score interpretations for this population of students and for ensuring that decisions made from the scores are fair for both ELs and non-ELs alike. Establishing validity of test scores for a given use is an ongoing practice, not a one-time event. It is important to continue to gather evidence of different types for different populations to make a robust argument that the scores are valid not only for a specific use, but for specific populations.

Limitations

One limitation of this study was that we were unable to disaggregate the effects of the supports by type of support offered. This is because 83% of ELs who retested with supports were offered more than one support (57% were offered three or four supports). This means that the impact of any given support would be confounded with the impact of other supports, and the sample sizes for students testing with any single support were very small. Extra time was the most common support offered, with 99% of ELs who received one or more supports receiving extra time. Extra time is required for ELs to effectively make use of other supports such as a word-to-word bilingual dictionary, which was approved for use for 77% of the ELs who retested with supports. Additional research should further investigate the impact of specific supports once adequate sample sizes are available.

Identification of ELs in this study was based on two criteria. ELs who tested without supports were identified by self-report, which may be problematic if a student misunderstood the question or if students who did not respond to the question were systematically different from students who did respond. ELs who tested with supports were identified based on ACT's accommodations system, which relies on accurate student information to match ACT score data to ACT accommodations data.

It is possible that some students in the sample were misidentified based on inaccurate or missing information that was used to match students across the two databases, although the impact of this is expected to be minimal. Additionally, while students are approved for accommodations, we do not have information about whether the accommodations were actually used; for example, a student may have been approved for a word-to-word bilingual dictionary but did not actually use it during the test.

Another limitation of this study was the large proportion of missing demographic and high school grades data, particularly for ELs who retested with supports. It is possible that students who provided this information may have differed systematically from those who did not provide the information, and potential bias should be considered when interpreting the results of this study.

There are many challenges to studying the performance of ELs. ELs are a very diverse population, differing with respect to their native language, culture, proficiency in English as well in as their native language, prior academic experiences, number of years in the US, parent education, socioeconomic status, and a plethora of other factors. Once in the US, differences in assessments; cut scores; policies for identifying and reclassifying Els; supports offered in classrooms; levels of mainstreamed, modified, or sheltered instruction; and other factors all contribute to a host of different experiences that impact both how much academic content they are exposed to and how quickly students gain proficiency in English (Abedi, 2001, 2008; Abedi, Hofstetter, & Lord, 2004; Linquanti, Cook, Bailey, & MacDonald, 2016). EL status is also temporary, as ELs are expected to gain English proficiency (assuming they are receiving appropriate support and instruction), thereby increasing the likelihood that EL students in the study sample have widely varying levels of English proficiency as students move in and out of the group identified as current ELs. All of these factors make it difficult to determine whether specific supports may be more or less effective for specific students, and this information was not available for the students in this study.

Notes

1. Additional information about EL testing supports and eligibility requirements can be found on ACT's website: http://www.act.org/content/act/en/products-and-services/the-act/registration/accommodations/policy-for-el-supports-documentation.html (ACT, 2021).

References

Abedi, J. (2001). *Assessment and accommodations for English language learners: Issues and recommendations* (CRESST, Policy Brief 4). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment, 8*(3), 231–257.

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice, 27*(3), 17–31.

Abedi, J. (2009). English Language Learners with disabilities: Classification, assessment, and accommodation issues. *Journal of Applied Testing Technology, 10*(2), 1–30.

Abedi, J., Courtney, M., & Leon, S. (2003). *Research-supported accommodation for English language learners in NAEP* (CSE Tech. Rep. No. 586). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Davis, CA: University of California, Davis, Smarter Balanced Assessment Consortium.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Tech. Rep. No. 536). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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), 1–28.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content based performance: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Acosta, B, Rivera, C. & Willner, L. S. (2008). *Best practices in state assessment for accommodating English language learners: A Delphi study*. Washington, DC: The George Washington University Center for Equity and Excellence in Education. Retrieved from: http://files.eric.ed.gov/fulltext/ED539759.pdf

ACT. (2019a). *ACT test translated directions for English learners*. Iowa City, IA: ACT. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/ACT-Translated-Test-Directions-All-Translations.pdf

ACT. (2019b). *The condition of college & career readiness 2019*. Iowa City, IA: ACT. Retrieved from http://www.act.org/content/dam/act/unsecured/documents/cccr-2019/National-CCCR-2019.pdf.

ACT. (2020). *The ACT technical manual*. Version 2020.1. Iowa City, IA: ACT. Retrieved from http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf.

ACT. (2021). Requesting English learner supports. Retrieved from http://www.act.org/content/act/en/products-and-services/the-act/registration/accommodations/policy-for-el-supports-documentation.html

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

Callahan, R. M., & Shifrer, D. (2016). Equitable access for secondary English learner students: Course taking as evidence of EL program effectiveness. *Educational Administration Quarterly, 52*(3), 463–496. DOI: 10.1177/0013161X16648190

Camara, W. J., & Allen, J. (2017). *Does testing date impact student scores on the ACT?* Iowa City, IA: ACT.

Cawthon, S., Ho, E., Patel, P., Potvin, D., & Trundt, K. (2009). Multiple constructs and effects of accommodations on accommodated test scores for students with disabilities. *Practical Assessment, Research & Evaluation, 14*(18). Retrieved from https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1224&context=pare

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. Retrieved from https://media.carnegie.org/filer_public/e3/7d/e37d44d4-b800-4d83-a723-078b21d59078/ccny_report_2006_ellassessments.pdf

Herman, J. L., & Abedi, J. (2004). *Issues in assessing English language learners' opportunity to learn mathematics*. CSE Report 633. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Johnson, A. (2019). A matter of time: Variations in high school course-taking by Years-as-EL subgroup. *Educational Evaluation and Policy Analysis, 41*(4), 461–482. DOI: 10.3102/0162373719867087.

Kieffer, M., Rivera, M., & Francis, D. (2012). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments Book 4*. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved from https://files.eric.ed.gov/fulltext/ED537635.pdf

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*(4), 285–307.

Linquanti, R., Cook, H. G., Bailey, A. L., & MacDonald, R. (2016). *Moving toward a more common definition of English learner: Collected guidance for states and multi-state assessment consortia*. Washington, DC: Council of Chief State School Officers.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classification based on test scores. *Journal of Educational Measurement, 32*(2), 179–197.

Lord, F. M. (1965). A strong true score theory with applications. *Psychometrika, 30*, 239–270.

Lovett, B. J. (2011). Extended time testing accommodations: What does the research say? *Communique: The Newspaper of the National Association of School Psychologists, 39*(8), 1 & 14–15.

Moore, J. L. (in press). *English learners who take the ACT with testing supports: An examination of performance, demographics, and contextual factors*. Iowa City, IA: ACT.

Moore, J. L., Li, T., & Lu, Y. (2020). *Reliability of English learners' test scores*. Iowa City, IA: ACT. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/R1754-el-score-reliability-2020-05.pdf

NCES. (2021). *The condition of education: English language learners in public schools.* Washington, DC: NCES. Retrieved from https://nces.ed.gov/programs/coe/indicator_cgf.asp.

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*(4), 248–260.

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*(3), 10–28.

Sanchez, E., & Buddin, R. (2015). *How accurate are self-reported high school courses, course grades, and grade point average?* **Iowa City, IA: ACT. http://www.act.org/content/dam/act/unsecured/documents/WP-2015-03.pdf**

Sugarman, J. (2019). *The unintended consequences for English learners of using the four-year graduation rate for school accountability*. **Washington, DC: MPI National Center on Immigrant Integration Policy.**

US Department of Education. (2016). *Non-regulatory guidance: English learners and Title III of the Elementary and Secondary Education Act (ESEA), as amended by the Every Student Succeeds Act (ESSA), appendix A*. **Washington, DC: US Department of Education. Retrieved from** *https://www2.ed.gov/policy/elsec/leg/essa/essatitleiiiguidenglishlearners92016.pdf*.

About the Authors

Joann L. Moore is a senior research scientist in Applied Research specializing in prediction of secondary and postsecondary outcomes from academic and non-cognitive factors.

Dongmei Li is a lead psychometrician at ACT specializing in test equating, scaling, and growth modeling.

Yang Lu is a former senior psychometrician at ACT specializing in multidimensional IRT, test equating, and educational measurement theories.

Acknowledgements

The authors would like to thank Krista Mattern and Joyce Schnieders for their comments on earlier drafts of this report.