Reliability of English Learners' Test Scores

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

The Every Student Succeeds Act requires that English Learners (ELs) are included in annual state testing (grades 3-8 and once in high school) and included in each state's accountability system disaggregated by subgroup to ensure that they receive the support they need to learn English, participate fully in their education experience, and graduate ready for college or career (US Department of Education, 2016). As more states are using the ACT® test as part of their accountability systems, one concern that educators and policymakers may have is whether the scores of English Learners (ELs) are valid and reliable indicators of their actual academic achievement level. This research brief will address the following research questions:

1. How does the reliability of ELs' ACT scores compare to that of non-ELs?

2. How does the reliability of ACT scores for ELs compare to the reliability of other standardized assessment scores?

3. How does classification consistency and differential item functioning analyses provide additional evaluative information about score validity?

Limited English proficiency can be a source of construct-irrelevant variance (measurement error), meaning that ELs' performance on a test may be negatively impacted because they have trouble comprehending the test content in the language in which the test is presented. As a result, their scores may not reflect their true ability level, particularly if the test has a high reading component.[1] This manifests in lower scores, as well as lower reliability estimates. Limited English proficiency can also be a source of construct-*relevant* variance if English proficiency is part of the construct being measured (e.g., English grammar), resulting in lower scores that do accurately reflect students' (lower) proficiency level.

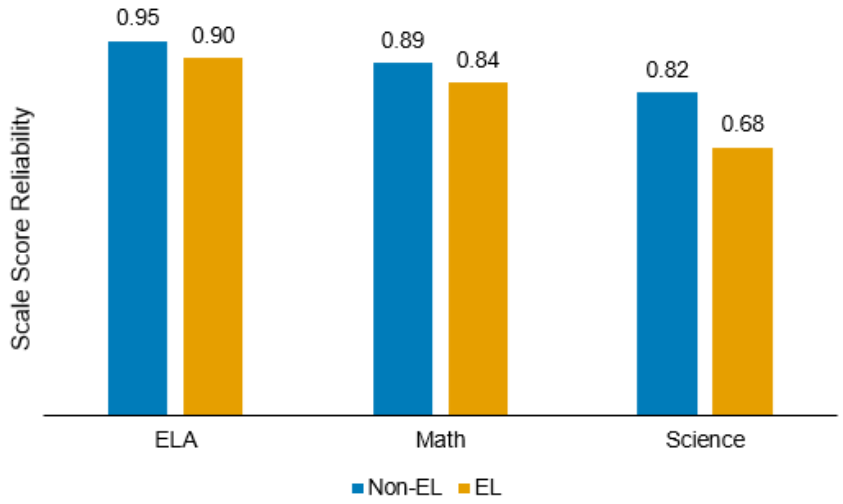
Reliability is a measure of the extent to which test scores are consistent across testing conditions, such as across different test items or upon retest. Cronbach's alpha is a common measure of internal consistency reliability (i.e., the extent to which students consistently respond to items sampled from the construct being measured).[2] A student who has mastered a construct should be able to consistently answer questions correctly, whereas a student who has not mastered the construct would be expected to consistently answer questions incorrectly. In contrast, an EL who knows the correct answer but is unable to comprehend the item content may produce an incorrect response that does not reflect their true knowledge.

Most of the research on the performance of ELs taking assessments in English has focused on average test scores, finding that ELs tend to score lower than non-ELs, particularly in content areas having a heavier language component (Abedi, 2002, 2003; Abedi, Leon, & Mirocha, 2003). However, several studies have found that the scores of ELs are also less reliable, have lower factor loadings, and have lower correlations with external measures than the scores of non-ELs (Abedi, 2002, 2009; Lakin & Lai, 2012). Score reliability may be lower for ELs with lower levels of English proficiency (Abedi, Leon, & Mirocha, 2003), and reliability estimates for ELs may be further attenuated due to restricted range of scores resulting from lower group performance (Lane & Leventhal, 2015). Additionally, while providing testing supports or accommodations may improve the scores of ELs, they typically do not completely eliminate the performance gap, and ELs who receive supports may have lower English proficiency levels than those who do not (Kieffer, Lesaux, Rivera, & Francis, 2009; Moore, Huang, Huh, Li, & Camara, 2018).

Reliability of ELs' ACT Scores Compared to Non-ELs

Figure 1 contains ACT scale score reliability estimates from a national sample of students (10,235 EL and 26,378 non-EL students) who took the ACT test as part of 2018 state and district testing.[3] Results are also presented in Table A1 of the Appendix. Across subject areas, score reliability for ELs was lower than score reliability for non-ELs. Small differences were found for English language arts (ELA) and Math (0.04-0.05), with a larger difference for science (0.14). Note that the ELA score presented here is not ACT's official ELA score, which is a composite based on the average of the English, reading, and writing section tests. The ELA score presented here is an average of the English and reading tests, excluding writing so that analyses could be conducted that require only multiple-choice items.

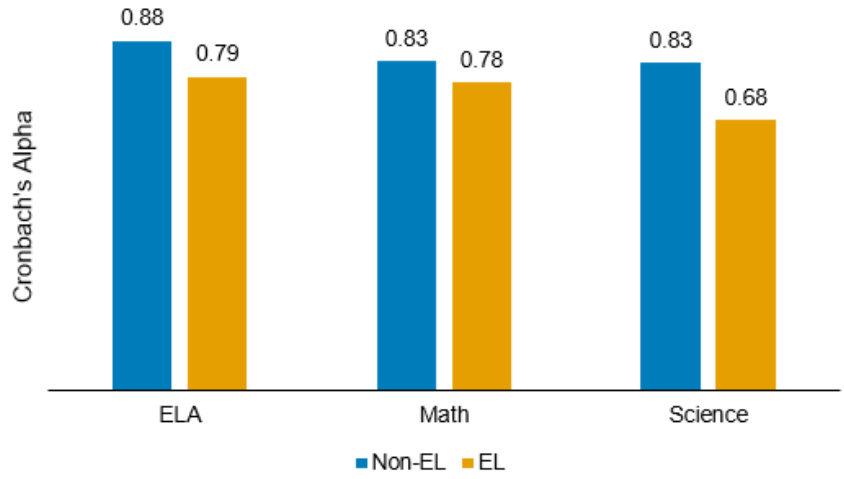
Figure 1. Scale Score Reliability Estimates for the ACT



Reliability of ACT Scores Compared to Other Standardized Assessments

Figure 2 contains reliability estimates from a summary of five studies that reported score reliability for ELs and non-ELs. A total of 64 comparisons were summarized across students in grades 2-11 and undergraduate and graduate applicants and across several assessments including the ITBS (Abedi, Leon, & Mirocha, 2003), NAEP items (Abedi, Lord, Hofstetter, & Baker, 2000), state assessments (Abedi, 2009; Young, Cho, Ling, Cline, Steinberg, & Stone, 2008), the Stanford 9 (Abedi, Leon, & Mirocha, 2003), and the SAT-Verbal and GRE-Verbal (Hale, Stansfield, & Duran, 1984). Results were averaged across ELA, language, reading, and verbal assessments into an overall ELA category (25 comparisons), across math assessments into an overall math category (34 comparisons), and across science assessments into an overall science category (5 comparisons).

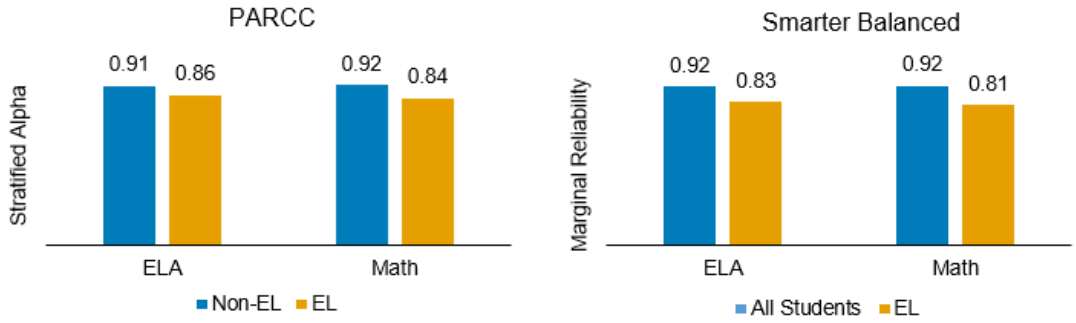
Figure 2. Average Cronbach's Alpha Reliability Estimates for EL and Non-EL Students Across Five Studies



Similar to the ACT results, on average, the reliability of ELs' scores were consistently lower than those of non-ELs; in fact, across the 64 comparisons, only three comparisons yielded estimates that were equal or higher for ELs—all in math. In general, and consistent with previous studies (Abedi, 2002; Abedi, Leon, & Mirocha, 2003), reliability gaps were smaller in math and smaller in earlier grade levels (see Appendix, Table A2). The larger reliability differences found for science assessments were consistent with findings for ACT science.

Figure 3 contains reliability estimates from the PARCC (2017) and Smarter Balanced Assessment Consortium (2016) technical manuals, by subject area, averaged across grades 3-11. PARCC reported stratified alpha estimates for EL and Non-EL students, and Smarter Balanced reported IRT-based marginal reliability estimates for EL students and for all students. Similar to previous research findings, average reliability estimates were lower for EL students as compared to non-EL students across subject areas and grade levels, and differences in reliability estimates were smaller at lower grade levels, while results were mixed for math (see Appendix, Tables A3 and A4).

Figure 3. Reliability Estimates for PARCC and Smarter Balanced Assessments



In summary, test score reliability in general is somewhat lower for ELs than for non-ELs. This is seen across many assessments, grade levels, and subject areas, and is more pronounced for students with lower levels of English proficiency and for content that requires a heavier reading component (Abedi, 2002, 2003; Abedi, Leon, & Mirocha, 2003). Lower reliability estimates were found for ELs taking the ACT, and the differences in reliability estimates were similar to, and in some cases smaller, than those found for other assessments, including other assessments that are being used for federal accountability purposes.

Classification Consistency and DIF

Classification consistency values were computed for ELA,[4] math, and science using the same examinees included in the reliability estimates (Figure 1). Classification for math and science was based on the ACT College Readiness Benchmarks, which are the scores at which a student has a 50% chance of earning a B or higher in first-year, credit bearing college courses. For the modified ELA score, which is the average of the ACT English and reading section tests, a cut score was derived using the same methodology as that used to develop the Benchmarks. Classification consistencies were calculated using the Livingston and Lewis (1995) method.

Table 1 presents a summary of the agreements between the operational test classifications—that is, the percentages of students who would be consistently classified in the same achievement levels on two equivalent administrations of the test. The agreement rate (percentage consistently classified) and Kappa index (agreement rate taking chance into account) were computed for each test score. The agreement rates were high (greater than 0.9) for both groups and were higher for ELs than for non-ELs. The Kappa statistics were all moderate to high and were lower for ELs in ELA and science and similar in mathematics.

Table 1. Consistency Indexes for Performance Levels

	EL		Non-EL	
Subject (Benchmark/cut score)	Agreement	Kappa	Agreement	Kappa
ELA (20)	0.960	0.806	0.923	0.838
Math (22)	0.974	0.803	0.935	0.809
Science (23)	0.948	0.536	0.894	0.687

Differential item functioning (DIF) analyses were also conducted using the same set of examinees. DIF can be described as a statistical difference between the probability of the specific population group (the "focal" group) getting the item right and the comparison population group (the "reference/base" group) getting the item right given that both groups have the same level of achievement with respect to the content being tested.[5]

Using the Mantel-Haenszel (MH) procedure, items with MH-D absolute values smaller than 1 were categorized as having negligible DIF, items with MH-D absolute values between 1 and 1.5 were flagged as having moderate DIF and items with MH-D absolute values of 1.5 or higher were flagged as having large DIF. Using the standardized difference in proportion-correct (STD) procedure, items were flagged when the values of STD were higher than 0.10. Table 2 shows the DIF analysis results based on the MH procedure, and Table 3 shows the DIF analysis results based on the STD procedure. No items were flagged for DIF.

Table 2. Summary of DIF Classifications with MH Procedure

Subject	Reference Group	Focal Group	N of Items	Flagged
English	Non-EL	EL	75	0
Math	Non-EL	EL	60	0
Reading	Non-EL	EL	40	0
Science	Non-EL	EL	40	0

Table 3. Summary of DIF Classifications with STD Procedure

Subject	Reference Group	Focal Group	N of Items	Flagged
English	Non-EL	EL	75	0
Math	Non-EL	EL	60	0
Reading	Non-EL	EL	40	0
Science	Non-EL	EL	40	0

In conclusion, limited English proficiency can be a source of measurement error when ELs are assessed in the English language such that students who have difficulty comprehending and responding accurately to test content may not be able to fully demonstrate their true achievement level. In addition, previous research has shown that ELs tend to perform at lower levels than their non-EL classmates when taking assessments in the English language, which can also attenuate reliability due to the restricted range of scores as compared to the full population. English proficiency is also a continuously moving target, and assuming ELs are receiving adequate support and instruction in learning English, they will eventually become proficient and be reclassified as former ELs. While there are challenges in measuring the proficiency of ELs, as this brief indicates, the reliability of ELs' scores on the ACT is comparable to that seen in other assessments and is sufficiently high that it does not by itself raise concerns about the validity of their scores. Additionally, we did not find evidence of DIF for ELs, and classification consistency analyses revealed similar agreement rates for ELs and non-ELs. These findings are encouraging and suggest that item characteristics are not introducing additional bias that would raise concerns about score validity for ELs.

Notes

1. In the fall of 2017, ACT began providing a limited number of testing supports to ELs in the United States taking the ACT test. The goal of these supports is to remove construct-irrelevant variance and allow ELs to more accurately demonstrate their true proficiency on the constructs being measured (Moore, Huang, Huh, Li, & Camara, 2018). The analyses presented in this paper include ELs who tested with (27%) and without (73%) supports. Future planned research will look specifically at the impact of these supports; this study is meant to address the reliability and validity of the scores of ELs in general.

2. Note that Cronbach's alpha estimates the reliability of number correct scores, whereas ACT reports a reliability estimate that is associated with scale scores, based on a four-parameter beta compound binomial model (Kolen, Hanson, & Brennan, 1992). Additional information can be found in the ACT Technical Manual (https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf).

3. EL and non-EL students were identified based on their 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" or were identified as EL based on ACT's Test Accessibility and Accommodations System were classified as EL, students who responded "No" were classified as non-EL, and students who responded "I prefer not to respond" were excluded from the analysis (approximately 66% of the sample).

4. Note that the ELA score presented here is the average of ACT English and reading, rather than ACT's ELA score, which is an average of English, reading, and writing.

5. The procedures used for the analysis include the standardized difference in proportion-correct (STD) procedure and the Mantel-Haenszel common odds-ratio (MH) procedure. For a description of these statistics and their performance overall in detecting DIF, see the ACT Research Report entitled *Performance of Three Conditional DIF Statistics in Detecting Differential Item Functioning on Simulated Tests* (Spray, 1989).

References

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

Abedi, J. (2003). Testing of English language learner students. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education*, (pp. 355-368). Washington, DC: American Psychological Association.

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

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of students' language background on content based assessment: Analyses of extant data* (CSE Technical Report No. 603). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://cresst.org/publications/cresst-publication-2975/.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3)*,* 16–26.

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). *Summaries of studies involving the Test of English as a Foreign Language, 1963-1982* (RR 84-3). Princeton, NJ: Educational Testing Service.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3)*,* 1168-1201.

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.

Lakin, J. M., & Lai, E. R. (2012). Multigroup generalizability analysis of verbal, quantitative, and nonverbal ability tests for culturally and linguistically diverse students. *Educational and Psychological Measurement, 72*(1), 139-158.

Lane, S., & Leventhal, B. (2015). Psychometric challenges in assessing English language learners and students with disabilities. *Review of Research in Education, 39*(1)*,* 165-214.

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.

Moore, J., Huang, C., Huh, N., Li, T., & Camara, W. (2018). *Testing supports for English learners: A literature review and preliminary ACT research findings*. Iowa City, IA: ACT. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/R1696-el-supports-2018-05.pdf.

PARCC. (2017). *Final technical report for 2016 administration*. Pearson. Retrieved from https://parcc-assessment.org/wp-content/uploads/2018/02/PARCC-2016-Tech-Report.pdf

Smarter Balanced Assessment Consortium (2016). *Smarter Balanced Assessment Consortium: 2014-15 technical report*. (Updated October 5, 2016). Los Angeles, CA: SBAC. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Spray, J. A. (1989). *Performance of Three Conditional DIF Statistics in Detecting Differential Item Functioning on Simulated Tests*. ACT Research Report Series 89-7. Iowa City, IA: ACT.

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.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13*(2-3), 170-192.

Appendix

Reliability Estimates by Subject Area and Grade Level Category

Table A1. Raw and Scale Score Reliability Estimates for EL and Non-EL Students, ACT

Subject Area	Non-EL (N = 26,378)	EL (N = 10,235)	Difference (Non-EL - EL)
	Cronbach's Alpha Raw Score Reliability Estimates		
ELA	0.95	0.91	0.04
Math	0.90	0.83	0.06
Science	0.85	0.74	0.11
	Four-Parameter Beta Binomial Scale Score Reliability Estimates		
ELA	0.95	0.90	0.04
Math	0.89	0.84	0.05
Science	0.82	0.68	0.14

Table A2. Average Cronbach's Alpha Reliability Estimates for EL and Non-EL Students across Five Studies Reporting Reliability

			Average Reliability		
Subject Area	Grade Levels	Number of Comparisons	Non-EL	EL	Difference (Non-EL - EL)
ELA	Grades 2-5	8	0.90	0.84	0.06
	Grades 6-8	8	0.86	0.76	0.11
	Grades 9-12	7	0.87	0.78	0.09
	Postsecondary	2	0.93	0.77	0.15
Math	Grades 2-5	13	0.84	0.81	0.03
	Grades 6-8	17	0.82	0.74	0.08
	Grades 9-12	4	0.88	0.84	0.04
Science	Grades 2-5	1	0.89	0.76	0.13
	Grades 6-8	1	0.88	0.73	0.15
	Grades 9-12	3	0.79	0.64	0.15

Table A3. Average Stratified Alpha Reliability Estimates for EL and Non-EL Students, PARCC Assessments

Subject Area	Grade Levels	Number of Comparisons	Average Reliability		
			Non-EL	EL	Difference (Non-EL - EL)
ELA	Grades 2-5	6	0.90	0.85	0.04
	Grades 6-8	6	0.92	0.87	0.05
	Grades 9-12	5	0.92	0.86	0.06
Math	Grades 2-5	6	0.93	0.90	0.03
	Grades 6-8	6	0.92	0.85	0.07
	Grades 9-12	7	0.91	0.78	0.13

Table A4. Average Marginal Reliability Estimates for EL and All Students, Smarter Balanced Assessments

Subject Area	Grade Levels	Number of Comparisons	Average Reliability		
			All Students	EL	Difference (All - EL)
ELA	Grades 2-5	3	0.92	0.86	0.06
	Grades 6-8	3	0.92	0.81	0.11
	Grades 9-12	1	0.92	0.80	0.12
Math	Grades 2-5	3	0.94	0.88	0.06
	Grades 6-8	3	0.92	0.78	0.14
	Grades 9-12	1	0.89	0.67	0.22

Joann Moore, PhD

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

Tianili Li, PhD

Tianli Li is a senior psychometrician in Assessment Transformation specializing in educational measurement theories and practices, item response theory, test for special population.

Yang Lu, PhD

Yang Lu is a senior psychometrician in the Assessment Transformation department. She specializes in multidimensional IRT, test equating, and educational measurement theories.