

Annual Technical Report for ACCESS for ELLsTM Paper English Language Proficiency Test

Series 603, 2024-2025 Administration
Annual Technical Report No. 21B

Part 3: References and Acknowledgments

Prepared by Center for Applied Linguistics
Language Assessment Division
Psychometrics and Quantitative Research Team
June 2026

References

- Allen, N. L., Carlson, J. E., & Zalanak, C. A. (1999). *The NAEP 1996 technical report*. National Center for Education Statistics.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.
- American Institutes of Research. (2018). *ELPA21 technical report, Part I—Summative assessment*.
- Andrich, D. A. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.
- Brennan, R. (2004). Linking with equivalent group or single group design (LEGS) (Version 2.0) [Computer software]. Center for Advanced Studies in Measurement and Assessment.
- Center for Applied Linguistics. (2016). *ACCESS for ELLs Series 400 listening and reading scale maintenance: Technical brief*.
- Center for Applied Linguistics. (2017). *ACCESS for ELLs 2.0 speaking and writing score scale reconstruction: Technical brief*.
- Center for Applied Linguistics. (2018). *Annual technical report for ACCESS for ELLs 2.0 Online English Language Proficiency Test, Series 401, 2016–2017 administration*.
- Center for Applied Linguistics. (2022). *The WIDA English Language Proficiency Assessment for Grades 1–12: test and item design plan for ACCESS for ELLs Paper and WIDA Screener Paper*.
- Cook, H. G., & MacGregor, D. (2017). *The ACCESS for ELLs 2.0 2016 Standard-setting study*. [Technical Report]. Board of Regents of the University of Wisconsin System.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Elementary and Secondary Education Act of 1965, amended 2015. 20 USC §1111(b)(1)(F); §1111(b)(2)(G) (2015).
- Engelhard, G., Jr., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge/Taylor & Francis Group.

- Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to proficiency levels: Guidelines for assessment design. *Applied Measurement in Education, 15*(3), 269–294.
- Every Student Succeeds Act (ESSA), 20 U.S.C. § 6301 (2015).
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Macmillan.
- George, D. & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 Update* (4th ed.) Allyn & Bacon.
- Gottlieb, M. (2004). *English language proficiency standards for English language learners in kindergarten through grade 12: Framework for large-scale state and classroom assessment*. WIDA Consortium.
- Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite scale scores*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education, 17*, 221–240.
- Kenyon, D. M. (2006). Development and field test of ACCESS for ELLs® (WIDA Consortium Technical Report No. 1). Center for Applied Linguistics.
- Kenyon, D. M., Ryu, J. R., & MacGregor, D. (2013). *Setting grade level cut scores for ACCESS for ELLs* [WIDA Technical Report No. 4]. Center for Applied Linguistics.
- Kim, A., Kondo, A., Blair, A., Mancilla, L., Chapman, & M., Wilmes, C. (2016). *Interpretation and use of K–12 language proficiency assessment score reports: Perspectives of educators and parents*. WCER Working Paper No. 2016-8. Board of Regents of the University of Wisconsin System.
- Kim, A., Chapman, & M., Kondo, A., & Wilmes, C. (2020). Examining the assessment literacy required for interpreting score reports: A focus on educators of K-12 English learners. *Language Testing, 37*(1) 54–75.
- Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement. *Journal of Educational Measurement, 29*, 285–307.
- Lance, C. E., Butts, M. M., & Michaels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods 9*(2), 202–220.
- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412–432.

- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.
- Linacre, J. M. (2006). Winsteps Rasch analysis (Version 5.2.4.0) [Computer software]. <http://www.winsteps.com>
- Livingston, S. A. (2018). *Test reliability—Basic concepts*. [Research Memorandum No. RM-18-01]. Educational Testing Service.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Lawrence Erlbaum Associates.
- MacGregor, D., Kenyon, D. M., Gibson, S., & Evans, E. (2009). *Development and field test of Kindergarten ACCESS for ELLs*. [WIDA Technical Report]. Board of Regents of the University of Wisconsin System.
- Mantel, N., & Haenszel, W. (1959). Statistical aspect of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Meyer, J. P. (2018). jMetrik [Computer software]. Retrieved from: <http://itemanalysis.com/jmetrik-download/>
- Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement*, 17(4), 151–363.
- National Center on Educational Outcomes. (2021). *Universal design of assessments*. https://nceo.info/Assessments/universal_design
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- Price, L. R., Lurie, A., Raju, N., Wilkins, C., & Zhu, J. (2006). Conditional standard errors of measurement for composite scores on the Wechsler Preschool and Primary Scale of Intelligence – Third Edition. *Psychological Reports*, 98(1), 237–252.
- Rudner, L. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice*, 20(1), 16–19.
- Sahakyan, N., (2020). *Generating alternate overall composite scale scores for English Learners with disabilities who are missing domain scores in the ACCESS for ELLs assessment*. [WIDA Technical Report.] Board of Regents of the University of Wisconsin System.
- U.S. Census Bureau. (n.d.) *About the topic of race*. Retrieved from: <https://www.census.gov/topics/population/race/about.html>

- U.S. Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*. Retrieved from <https://www.ed.gov/sites/ed/files/2020/07/assessmentpeerreview.pdf>
- Waller, N. G. (n.d.). EZDIF: A computer program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel and logistic regression procedures [Computer software]. University of California – Davis.
- WIDA. (2007). *English language proficiency standards and resource guide, 2007 edition, prekindergarten through grade 12*. Board of Regents of the University of Wisconsin System.
- WIDA. (2012). *2012 amplification of the English language development standards kindergarten–grade 12*. Board of Regents of the University of Wisconsin System.
- WIDA. (2021). *WIDA's Equipercentile Method and Procedure for Comparability Analysis*. Board of Regents of the University of Wisconsin System.
<https://sea.wida.us/documents/widas-equipercenile-method-and-procedure-comparability-analysis>
- WIDA. (2025). *ACCESS for ELLs Interpretive Guide for Score Reports*. Board of Regents of the University of Wisconsin System.
- WIDA. (n.d.) ACCESS for ELLs Test Practice and Sample Items.
<https://wida.wisc.edu/assess/access/preparing-students/practice>
- WIDA. (n.d.). WIDA Assessment Use Argument. [Unpublished internal report.]
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. MESA Press.
- Young, M. J., & Yoon, B. (1998, April). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment* [CSE Technical Report 475]. University of California-Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Erlbaum.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30, 233–251.

Acknowledgments

We would like to extend our appreciation to the many CAL and WIDA staff members who have supported this work.

From CAL:

Tanya Bitterman, M.A.
Sofia Buitrago, MS.
Seong Eun Hong, Ph.D.
Michele Kawood, M.S.Ed.
Justin Kelly, Ph.D.
Reshmi Kumpakha, M.A.
Hyeonseong Lee, M.A.
Samantha Musser, M.A.
Aubrey Sahouria
Yu Lan Su, Ph.D.
Seyfullah Tingir, Ph.D.
Jasmine Tsai, M.Ed.
Yamei Wang, Ph.D.
Frank Wucinski, M.A.
Shu Jing Yen, Ph.D.

From WIDA:

Kyoungwon Bishop, Ph.D.
Hacer Karamese, Ph. D.
John Lorge