Bibliography

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014) (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (pp. 508–600). Washington, DC: American Council on Education.

Center for Applied Special Technology (CAST). (2018). *Universal design for learning (UDL) guidelines - version 2.2.* Retrieved from http://udlguidelines.cast.org.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Publishing Company.

Davies, S., O'Malley, K., & Wu, B. (2007, April). *Establishing measurement equivalence of transadapted reading and mathematics tests*. Paper presented at the 2007 annual meeting of the American Educational Research Association, Chicago.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7(1), 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Ferrara, S., Lewis, D., Mercado, R., D'Brot, J., Barth, J., & Egan, K. (2011, April). *A method for setting benchmarked performance standards: Workshop procedures, panelist judgments, and empirical results*. Paper presented at the 2011 annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41–56.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.

Kane, M. T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Linacre, J. M. (2018). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.

O'Malley, K., Keng, L., & Miles, J. (2012). Using validity evidence to set performance standards. In G.J. Cizek (Ed.), *Setting performance standards* (pp. 301–322). New York: Routledge.

Petersen, N. S. (1987, September 25). *DIF procedures for use in statistical analysis* [ETS internal memorandum].

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational Measurement* (pp. 221–262). New York: Macmillan.

Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 342–364). New York: Routledge.

Rasch, G. (1966). An individualistic approach to item analysis. In P. Lazarsfeld & N. W. Henry (Eds.), *Readings in Mathematical Social Science* (pp. 89–107). Chicago: Science Research Associates.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(14). Available online: http://pareonline.net/getvn.asp?v=7&n=14.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, 10(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13.

Schafer, W. D., Wang, J., & Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. In R. W. Lissitz (Ed.) *The concept of validity: revisions, new directions, and applications* (pp. 173–193). Charlotte, NC: Information Age.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and practice*, 16(2), 5–8, 13, 24.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Wingersky, M.S., & Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, 8(3), 347-364.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97–116.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: MESA Press.

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.