Bibliography

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *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 (2002). *Meeting Diverse Learner Needs Through Universal Design for Learning.* Retrieved June 4, 2003 from http://www.cast.org/udl/meetingdiverselearnerneeds2519.cfm.

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 AERA, Chicago, IL.

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, NY: 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 meetings of the NCME, 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 NCME & the American Council on Education.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. New York, NY: Springer.

Linacre, J. M. (2001). *A User's Guide to WINSTEPS: Rasch-Model Computer Program*. Chicago, IL: MESA Press.

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, NY: Routledge.

Petersen, N. S. (1987, September 25). *DIF Procedures for Use in Statistical Analysis.* Educational Testing Service 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, NY: Macmillan.

Phillips, G. W. (2012). "The Benchmark Method of Standard Setting." In G. Cizek (Ed.), *Setting Performance Standards* (pp. 342–364). New York, NY: Routledge.

Rasch, G. (1966). "An Individualistic Approach to Item Analysis." In P. Lazarfeld & N. W. Henry (Eds.), *Readings in Mathematical Social Science* (pp. 89–107). Chicago, IL: 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* (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, NY: 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 NCME, 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.