

References

- Algina, J., & Noe, M. J. (1978). A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. *Journal of Educational Measurement, 15*, 101-110.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.
- Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: ACT Publications.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.
- Committee of AERA, APA, & NCME. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.
- Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—or should we? *Psychological Bulletin, 74*, 68-80.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.
- de Gruijter, D. N. M., & van der Kamp, L. J. Th. (1984). *Statistical models in psychological and educational testing*. Lisse: Swets & Zeitlinger.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.
- Fleishman, J., & Benson, J. (1987). Using LISREL to evaluate measurement models and scale reliability. *Educational and Psychological Measurement, 47*, 925-939.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.
- Guttman, L. (1953). A special review of Harold Gulliksen, *Theory of mental tests*. *Psychometrika, 18*, 123-130.
- Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*, 159-170.

- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6, 373-378.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13, 253-264.
- Huynh, H. (1978). Reliability of multiple classifications. *Psychometrika*, 43, 317-325.
- Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 335-355). Washington, DC: American Council on Education.
- Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7 user's reference guide*. Mooresville, IL: Scientific Software.
- Keeping, E. S. (1962). *Introduction to statistical inference*. Princeton, NJ: Van Nostrand.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160.
- Linn, R. L., & Werts, C. E. (1979). Covariance structures and their analysis. In R. E. Traub (Ed.), *New directions for testing and measurement: Methodological developments* (No. 4, pp. 53-73). San Francisco: Jossey-Bass.
- Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, 16, 421-437.
- Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement*, 18, 437-454.
- Lord, F. M. (1959). An approach to mental test theory. *Psychometrika*, 24, 283-302.
- Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Manning, W. H., & DuBois, P. H. (1962). Correlational methods in research on human learning. *Perceptual and Motor Skills*, 15, 287-321.
- Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York: Freeman.
- Merriam-Webster's Collegiate Dictionary* (10th ed.). (1993). Springfield, MA: Merriam-Webster.
- Millman, J. (1979). Reliability and validity of criterion-referenced test scores. In R. Traub (Ed.), *New directions for testing and measurement: Methodological developments* (No. 4, pp. 75-92). San Francisco: Jossey-Bass.
- MINITAB, Inc. (1989). *MINITAB statistical software* (Release 7). State College, PA: Author.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1-18.
- Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1-13.
- Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6, 1-9.
- Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 90, 726-748.
- Rogosa, D. R., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology*, 76, 1000-1027.

- Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, 20, 335-343.
- Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50, 203-228.
- Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split halves. *Harvard Educational Review*, 9, 99-103.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161-169.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295.
- Spray, J. A., & Welch, C. J. (1990). Estimation of classification consistency when the probability of a correct response varies. *Journal of Educational Measurement*, 27, 15-25.
- Stanley, J. C. (1967). General and special formulas for reliability of differences. *Journal of Educational Measurement*, 4, 249-252.
- Stevens, S. S. (1948). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: John Wiley.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13, 265-276.
- Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 15, 111-116.
- Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 11, 263-267.
- Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 12, 87-98.
- Traub, R. E., & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement*, 4, 517-545.
- Tucker, L. J., Damarin, F., & Messick, S. (1966). A base-free measure of change. *Psychometrika*, 31, 457-473.
- Willett, J. B. (1988-1989). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15, pp. 345-422). Washington, DC: American Educational Research Association.
- Woodruff, D. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement*, 27, 191-208.
- Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.