Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 53

Generalizability Theory References: The First Sixty Years

Robert L. Brennan[1]
and
Jui-Teng Ray Liao[2]

February 1, 2020

[1]Robert L. Brennan is E. F. Lindquist Professor Emeritus in Measurement and Testing and Founding Director, Center for Advanced Studies in Measurement and Assessment (CASMA).

[2]Research Assistant, Center for Advanced Studies in Measurement and Assessment (CASMA).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: https://education.uiowa.edu/casma

Most researchers associate the beginning of Generalizability (G) theory with Cronbach, Gleser, Nanda, and Rajaratnam (1972). That monograph, however, was the result of research conducted by the authors primarily in the 1960s. For that reason, somewhat arbitrarily, we view G theory as having an approximate 60 year history, although it could be argued that the genesis of G theory extends back to the 1920s.

This report provides a list of over 300 references that relate to G theory. These references are, in a sense, a supplement to Brennan (2020) who uses a subset of these references to support and illustrate his perspectives on the history of G theory. Because the G theory framework is so broad, the boundaries of the theory are sometimes indistinct. Some might argue that certain references provided here are only marginally relevant; others might argue that this report does not include some/many relevant references. The authors invite suggestions about additional references that might be included in a revised version of this report. (Please provide suggested additional references in APA style, at least approximately.)

Separately alphabetized sets of references are provided for the 20th and 21st centuries, which is roughly consistent with the structure of Brennan (2020). (The 21st century references begin on page 14.) Note that the report does not include all references in Cronbach et al. (1972) or Brennan (2001b). Also, with some exceptions, presentations at meetings of professional associations are not included. The inclusion of a reference here does not imply any evaluative judgment of its content

20th Century

Algina, J. (1989). Elements of classical reliability theory and generalizability theory. *Advances in Social Science Methodology*, *1*, 137–169.

Allal, L. (1988). Generalizability theory. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement* (pp. 272–277). New York: Pergamon.

Allal, L. (1990). Generalizability theory. In H. J. Walberg, & G. D. Haertel (Eds.), *The international encyclopedia of educational evaluation* (pp. 274-279). Oxford, England: Pergamon.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Arteaga, C., Jeyaratnam, S., & Graybill, F. A. (1982). Confidence intervals for proportions of total variance in the two-way cross component of variance model. *Communications in Statistics: Theory and Methods*, *11*, 1643–1658.

Bachman, L. F., Lynch, B. K., & Mason, M. (1994). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*, 239–257.

Bell, J. F. (1985). Generalizability theory: The software problem. *Journal of Educational Statistics, 10*, 19–29.

Bell, J. F. (1986). Simultaneous confidence intervals for the linear functions of expected mean squares used in generalizability theory. *Journal of Educational Statistics*, *11*, 197–205.

Betebenner, D. W. (1998, April). *Improved confidence interval estimation for variance components and error variances in generalizability theory.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Bock, R. D. (1975).*Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Boodoo, G. M. (1982). On describing an incidence sample. *Journal of Educational Statistics*, *7*(4), 311–331.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

Brennan, R. L. (1977a). *Generalizability analyses: Principles and procedures.* (ACT Technical Bulletin No. 26). Iowa City, IA: ACT, Inc. (Revised August 1978).

Brennan, R. L. (1977b). *KR-21 and lower limits of an index of dependability for mastery tests.* (ACT Technical Bulletin No. 27). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1978a). *Extensions of generalizability theory to domain-referenced testing.* (ACT Technical Bulletin No. 30). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1978b). *Algorithms, procedures, and variance components for analysis of variance.* (ACT Technical Bulletin No. 31). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1979a). *Some applications of generalizability theory to the dependability of domain-referenced tests.* (ACT Technical Bulletin No. 32). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1979b). *Handbook for Gapid: A fortran IV computer program for generalizability analyses with single-facet designs.* (ACT Technical Bulletin No. 34). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1980). *Applications of generalizability theory.* In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art (pp. 186-232). Baltimore: The Johns Hopkins University Press.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: ACT, Inc.

Brennan, R. L. (1984a). *Some statistical issues in generalizability theory.* (ACT Technical Bulletin No. 46). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1984b). Estimating the dependability of the scores. In R. A.

Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 292–334). Baltimore: The Johns Hopkins University Press.

Brennan, R. L. (1989). Proof of equivalence of estimates of absolute error variance and average of Feldt's individual-level error variances in table of specifications model. In R. L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+.* (p. 91). Iowa City, IA: American College Testing.

Brennan, R. L. (1992a). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1992b). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

Brennan, R. L. (1994). Variance components in generalizability theory. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 175–207). New York: Plenum.

Brennan, R. L. (1995a). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, *14*, 385–396.

Brennan, R. L. (1995b). Standard setting from the perspective of generalizability theory. In *Proceedings of the joint conference on standard setting for large-scale assessments* (Volume II). Washington, DC: National Center for Education Statistics and National Assessment Governing Board.

Brennan, R. L. (1996a). *Conditional standard errors of measurement in generalizability theory* (Iowa Testing Programs Occasional Paper No. 40). Iowa City, IA: Iowa Testing Programs, University of Iowa.

Brennan, R. L. (1996b). Generalizability of performance assessments. In G. W. Phillips (Ed.). *Technical issues in performance assessments.* Washington, DC: National Center for Education Statistics.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice*, *16*(4), 14–20.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, *22*, 307–331.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement*, *55*, 157–176.

Brennan, R. L., Harris, D. J., & Hanson, B. A. (1987). *The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts* (American College Testing Research Report No. 87-7). Iowa City, IA: ACT, Inc.

Brennan, R. L., Jarjoura, D., & Deaton, E. L. (1980). *Some issues concerning the estimation and interpretation of variance components in generalizability theory.* (ACT Technical Bulletin No. 36). Iowa City, IA: ACT, Inc.

Brennan, R. L., & Lee, W. C. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Edu-*

cational and Psychological Measurement, 59(1), 5–24.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9–12.

Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*, 277–289.

Brennan, R. L., & Kane, M. T. (1977b). Signal/noise ratios for domain-referenced tests. *Psychometrika*, *42*, 609–625.

Brennan, R. L., & Kane, M. T. (1979). Generalizability theory: A review. In R. E. Traub (Ed.), *New directions for testing and measurement: Methodological developments* (No.4) (pp. 33–51). San Francisco, CA: Jossey-Bass.

Brennan, R. L., & Kane, M. T. (1985). *Generalizability theory: Analyzing measurement error in the assessment of artistic products and performances*. (ACT Technical Bulletin No. 49). Iowa City, IA: ACT, Inc.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, *4*, 219–240.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*(2), 217–238.

Burdick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components*. New York: Dekker.

Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *The marks of examiners*. London: The Macmillan Company.

Butterfield, P. S., Mazzaferri, E. L., & Sachs, L. A. (1987). Nurses as evaluators of the humanistic behavior of internal medicine residents. *Journal of Medical Education*, *62*, 842–849.

Cardinet, J., & Tourneur, Y. (1985). *Assurer la measure*. New York: Peter Lang.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, *13*, 119–135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, *18*, 183–204.

Chambers, D. W., & Loos, L. (1997). Analyzing the sources of unreliability in fixed prosthodontics mock board examinations. *Journal of Dental Education*, *61*, 346–353.

Clauser, B. E., Clyman, S. G., & Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement, 36*(1), 29–45.

Collins, J. D. (1970). *Jackknifing generalizability*. Unpublished doctoral dissertation, University of Colorado, Boulder.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, *27*, 907–949.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin No. 43). Iowa City, IA: ACT, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt.

Cronbach, L. J. (1947).

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 292–334.

Cronbach, L. J. (1976). On the design of educational measures. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 199–208). New York: Wiley.

Cronbach, L. J. (1989). Lee J. Cronbach. In G. Lindzey (Ed.), *A history of psychology in autobiography* (Vol. VIII). Stanford, CA: Stanford University Press.

Cronbach, L. J. (1991). Methodological studies-A personal retrospective. In R. E. Snow., & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 385–400). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, *24*, 467–480.

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: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995, Summer). Generalizability analysis for educational assessments (Evaluation Comment). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Cronbach, L.J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373–399.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137–163.

Cronbach, L. J., Schönemann, P., & McKie, T. D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, *25*, 291–312.

Crooks, T. J., & Kane, M. T. (1981). The generalizability of student ratings of instructors: Item specificity and section effects. *Research in Higher*

Education, *15*, 305–313.

Crowley, S. L., Thompson, B., & Worchel, F. (1994). The Children's Depression Inventory: A comparison of generalizability and classical test theory analyses. *Educational and Psychological Measurement*, *54*, 705–713.

Crump, S. L. (1946). The estimation of variance components in analysis of variance. *Biometrics Bulletin, 2*, 7-11.

Demorest, M. E., & Bernstein, L. E. (1993). Applications of generalizability theory to measurement of individual differences in speech perception. *Journal of the Academy of Rehabilitative Audiology*, *26*, 39–50.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*, 289–303.

Dunnette, M. D., & Hoggatt, A. C. (1957). Deriving a composite score from several measures of the same attribute. *Educational and Psychological Measurement, 17*, 423–434.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, *16*, 407–424.

Eisenhart, C. (1947). The assumptions underlying analysis of variance. *Biometrics, 3*, 1-21.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia: SIAM.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*, 54–77.

Erlich, O.,& Borich, C. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. *Journal of Educational Measurement, 16*, 11-18.

Erlich, O.,& Shavelson, R. J. (1976). *Application of generalizability theory to the study of teaching* (Technical Report No. 76-9-1). Beginning Teacher Evaluation Study, Far West Laboratory, San Francisco.

Feldt, L. S. (1965). The approximate sampling distribution of Kuder- Richardson reliability coefficient twenty. *Psychometrika*, *30*, 357–370.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 105–146). New York: American Council on Education and Macmillan.

Feldt, L. S., & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33,* 141–156.

Finn, A., & Kayandé, U. (1997). Reliability assessment and optimization of marketing measurement. *Journal of Marketing Research*, *34*, May, 262–275.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver & Bond.

Gao, X., Brennan, R. L.,& Shavelson, R. J. (1994, April). *Estimating generalizability of matrix-sampled science performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, *15*, 1–14.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*, 395–418.

Graybill, F. A. (1976). *Theory and application of the linear model.* North Scituate, MA: Duxbury Press.

Graybill, F. A., & Wang, C. M. (1980). Confidence intervals on nonnegative linear combinations of variances. *Journal of the American Statistical Association*, *75*, 869–873.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley. [Reprinted by Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.]

Hartley, H. O. (1967). Expectations, variances, and covariances of ANOVA mean squares by 'synthesis.' *Biometrics, 23*, 105–114, and Corrigenda, 853.

Hartley, H. O., Rao, J. N. K., & LaMotte, L. R. (1978). A simple 'synthesis'-based method of variance component estimation. *Biometrics, 34*, 233–242.

Hartman, B. W., Fuqua, D. R., & Jenkins, S. J. (1988). Multivariate generalizability analysis of three measures of career indecision. *Educational and Psychological Measurement*, *48*, 61–68.

Hatch, J. P., Prihoda, T. J., & Moore, P. J. (1992). The application of generalizability theory to surface electromyographic measurements during psychophysiological stress testing: How many measurements are needed? *Biofeedback and Self Regulation*, *17*, 17–39.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, *9*, 227–252.

Hoover, H. D., & Bray, G. B. (1995, April). *The research and development phrase: Can a performance assessment be cost-effective?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*, 153–160.

Huynh, H. (1977, April). *Estimation of the KR20 reliability coefficient when data are incomplete.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Jarjoura, D. (1983). Best linear prediction of composite universe scores. *Psychometrika*, *48*(4), 525–539.

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement, 10*, 175–186.

Jarjoura, D., & Brennan, R. L. (1981, January). *Three variance components models for some measurement procedures in which unequal numbers of items fall into discrete categories* (American College Testing Technical Bulletin No. 37). Iowa City, Iowa: ACT, Inc.

Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement*, *6*, 161–171.

Jarjoura, D., & Brennan, R. L. (1983). Multivariate generalizability models for tests developed according to a table of specifications. In L. J. Fyans (Ed.), *New directions for testing and measurement: Generalizability theory: Inferences and practical applications* (No.18) (pp. 83–101). San Francisco, CA: Jossey-Bass.

Jarjoura, D., Hartman-Stein, P., Speight, J., & Reuter, J. (1999). Reliability and construct validity of scores on the behavioral competence inventory: A measure of adaptive functioning. *Educational and Psychological Measurement, 59*(5), 855–865.

Joe, G. W., and Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika*, *41*(2), 205–217.

Johnson, S., & Bell, J. F. (1985). Evaluating and predicting survey efficiency using generalizability theory. *Journal of Educational Measurement*, *22*, 107–119.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Associates.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125–160.

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education*, *9*, 355–379.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, *47*, 267–292.

Kane, M. T., Crooks, T. J., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement*, *13*, 171–183.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics (4th ed., Vol. 1)*. New York: Macmillan.

Khuri, A. I. (1981). Simultaneous confidence intervals for functions of vari-

ance components in random models. *Journal of the American Statistical Association*, *76*, 878–885.

Klipstein-Grobusch, K., Georg, T., & Boeing, H. (1997). Interviewer variability in anthropometric measurements and estimates of body composition. *International Journal of Epidemiology*, *26*(Suppl. 1), 174–180.

Koch, G. G. (1968). Some further remarks concerning "A general approach to the estimation of variance components." *Technometrics, 10*, 551–558.

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9*, 209–223.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices.* New York: Springer-Verlag.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*, 285-307.

Kolen M. J.,& Jarjoura, D. (1984). Item profile analysis for tests developed according to a table of specifications. *Applied Psychological Measurement, 8*, 219-230.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29,* 285–307.

Kolen, M. J. & Harris, D. J. (1987, April). *A multivariate test theory model based on item response theory and generalizability theory.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Knight, R. G., Ross, R. A., Collins, J. I., & Parmenter, S. A. (1985). Some norms, reliability and preliminary validity data for an S-R inventory of anger: The Subjective Anger Scale (SAS). *Personality and Individual Differences*, *6*, 331–339.

Kreiter, C. D., Brennan, R. L., & Lee, W. (1998). A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performance. *Academic Medicine*, *73*, 1294–1298.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–160.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33*, 71–92.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37,* 1–20.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2005). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19* 4, 9–15.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*, 237–255.

Leucht, R. M., & Smith, P. L. (1989, April). *The effects of bootstrapping strategies on the estimation of variance components*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton-Mifflin.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5–8, 15.

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, CA: Jossey-Bass.

Llabre, M. M., Ironson, G. H., Spitzer, S. B., Gellman, M. D., Weidler, D. J., & Schneiderman, N. (1988). How many blood pressure measurements are enough?: An application of generalizability theory to the study of blood pressure reliability. *Psychophysiology*, *25*, 97–106.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*, 143–155.

Longford, N. T. (1995). *Models for uncertainty in educational testing.* New York: Springer-Verlag.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325–336.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, *17*, 510–521.

Lord, F. M. (1959). Tests of the same length do have the same standard error of measurement? *Educational and Psychological Measurement*, *19*, 233–239.

Lord, F. M. (1962). Test reliability: A correction. *Educational and Psychological Measurement, 22*, 511-512.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loveland, E. H. (1952). *Measurement of factors affecting test-retest reliability.* Unpublished doctoral dissertation, University of Tennessee.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Marcoulides, G. A. (1998). Applied generalizability theory models. In G. A. Marcoulides (Ed.), *Modern methods for business research*. Mahwah, NJ: Erlbaum.

Marcoulides, G. A., & Goldstein, Z. (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement*, *50*, 761–768.

Marcoulides, G. A., & Goldstein, Z. (1992). The optimization of multivariate generalizability studies with budget constraints. *Educational and Psychological Measurement*, *52*, 301–308.

Miller, T. B., & Kane, M. T. (2001, April). *The precision of change scores under absolute and relative interpretations.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement*, *24*, 56–64.

Nuβbaum, A. (1984). Multivariate generalizability theory in educational measurement: An empirical study. *Applied Psychological Measurement*, *8*(2), 219–230.

Oppliger, R. A., & Spray, J. A. (1987). Skinfold measurement variability in body density prediction. *Research Quarterly for Exercise and Sport*, *58*, 178–183.

Othman, A. R. (1995). *Examining task sampling variability in science performance assessments.* Unpublished doctoral dissertation, University of California, Santa Barbara.

Quenouille, M. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society B*, *11*, 18–24.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39–56.

Rentz, J. O. (1987). Generalizability theory: A comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research*, *24*, (February), 19-28.

Rogosa, D.,& Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics, 3*, 157-252.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, *30*, 41–53.

Saab, P. G., Llabre, M. M., Hurwitz, B. E., Frame, C. A., Reineke, L. J., Fins, A. I., McCalla, J., Cieply, L. K.,& Schneiderman, N. (1992). Myocardial and peripheral vascular responses to behavioral challenges and their stability in black and white Americans. *Psychophysiology, 29*, 384-397.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*, 309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Searle, S. R. (1971). *Linear models.* New York: Wiley.

Searle, S. R. (1974). Prediction, mixed models, and variance components. In F. Proschan & R. J. Sterfling (Eds.), *Reliability and biometry.* Philadelphia: SIAM.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York: Wiley.

Shao, J., & Tu, D. (1995). *The jackknife and the bootstrap.* New York: Springer-Verlag.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215–232.

Shavelson, R. J., Baxter, G. P.,& Pine, J. (1991). Performance assessments in science. *Applied Measurement in Education, 4*, 347-362.

Shavelson, R. J., Baxter, G. P.,& Pine, J. (1992). Performance assessments: The rhetoric and reality. *Educational Researcher, 21*(4), 22-27.

Shavelson, R. J., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*, 553–611.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133–166.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shavelson, R. J., & Webb, N. M. (1992). Generalizability theory. In M. C. Alkin (Ed.), *Encyclopedia of Educational Research* (Vol. 2) (pp. 538–543). New York: Macmillan.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *6*, 922–932.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Sirotnik, K., & Wellington, R. (1977). Incidence sampling: An integrated theory for "matrix sampling." *Journal of Educational Measurement*, *14*, 343–399.

Smith, P. L. (1978). Sampling errors of variance components in small sample generalizability studies. *Journal of Educational Statistics*, *3*, 319–346.

Smith, P. L. (1982). A confidence interval approach for variance component estimates in the context of generalizability theory. *Educational and Psychological Measurement*, *42*, 459–466.

Thompson, B., & Melancon, J. G. (1987). Measurement characteristics of the Group Embedded Figures Test. *Educational and Psychological Measurement*, *47*, 765–772.

Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S., & Lu, T. C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computational Simulation*,

35, 135–143.

Tobar, D. A., Stegner, A. J., & Kane, M. T. (1999). The use of generalizability theory in examining the dependability of score on the Profile of Mood States. *Measurement in Physical Education and Exercise Science*, *3*, 141–156.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Ulrich, D. A., Riggen, K. J., Ozmun, J. C., Screws, D. P., & Cleland, F. E. (1989). Assessing movement control in children with mental retardation: A generalizability analysis of observers. *American Journal of Mental Retardation*, *94*, 170–176.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*, 1–21.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*, 1–14.

Wang, M. C., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*, 663–705.

Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of General Educational Development ratings. *Journal of Educational Measurement*, *18*, 13–22.

Webb, N. M., Shavelson, R. J., and Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans (Ed.), *New Directions in Testing and Measurement: Generalizability Theory*, (No. 18), 67–82. San Francisco, CA: Jossey-Bass.

Wiley, E. W. (2000). *Bootstrap strategies for variance component estimation: Theoretical and empirical results*. Unpublished doctoral dissertation, Stanford.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3*, 23–40.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Wohlgemuth, W. K., Edinger, J. D., Fins, A. I., & Sullivan, R. J. (1999). How many nights are enough? The short-term stability of sleep parameters in elderly insomniacs. *Psychophysiology*, 36, 233–244.

Wong, S. P., & McGraw, K. O. (1999). Confidence intervals and F tests for intraclass correlations based on three-way random effects models. *Educational and Psychological Measurement, 59*(2), 270–288.

21st Century

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement, 74*(5), 795–808.

Attali, Y. (2010). An analysis of variance approach for the estimation of response time distributions in tests. *Journal of Educational Measurement, 47*(4), 458–470.

Balogh, J., Bernstein, J., Cheng, J., Van Moere, A., Townshend, B., & Suzuki, M. (2012). Validation of automated scoring of oral reading. *Educational and Psychological Measurement, 72*(3), 435–452.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing writing, 12*(2), 86–107.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). *The information in multiple ratings. Applied Psychological Measurement, 26*, 364–375.

Block and Norman (2015)

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83–100.

Brennan, R. L. (2000a). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5–10.

Brennan, R. L. (2000b) Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*, 339–353.

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295-317.

Brennan, R. L. (2001b). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2001c). *Manual for mGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa. [Computer software and manual.] (Retrieved from https://education.uiowa.edu/casma-computer-programs)

Brennan, R. L. (2001d). *Manual for urGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa. [Computer software and manual.] (Retrieved from https://education.uiowa.edu/casma-computer-programs)

Brennan, R. L. (2003). Coefficients and indices in generalizability theory (CASMA Research Report No. 1). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Retrieved from https://education.uiowa.edu/casma)

Brennan, R. L. (2004). Generalizability theory. In M. S. Lewis-Beck, A. E. Bryman, & T. F. Liao (Eds.), *The SAGE encyclopedia of social science research methods* (Vol. 2), 418–420. Thousand Oaks, CA: SAGE.

Brennan, R. L. (2006a). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1-16). Westport, CT: American Council on Education/Praeger.

Brennan, R. L. (2006b, November). Unbiased estimates of variance components with bootstrap procedures: Detailed results. (CASMA Research Report No. 21). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
(Retrieved from https://education.uiowa.edu/casma)

Brennan, R. L. (2007a). Integration of models. In C. Rao and S. Sinharey (Eds.) *Handbook of statistics: Psychometrics* (Vol. 26) (pp. 1095–1098). Amsterdam: Elsevier.

Brennan, R. L. (2007b). Unbiased estimates of variance components with bootstrap procedures. *Educational and Psychological Measurement, 67*(5), 784–803.

Brennan, R. L. (2009, February). *Nominal weights in multivariate generalizability theory.* (CASMA Technical Note No. 4). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Retrieved from https://education.uiowa.edu/casma)

Brennan, R. L. (2010a). Generalizability theory. In P. Peterson, E. Baker, & B. McGaw (Eds.) *International Encyclopedia of Education* (3rd ed.), vol. 4, 61–68.

Brennan, R. L. (2010b). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1–21.

Brennan, R. L. (2011, September). *Using Generalizability Theory to Address Reliability Issues for PARCC Assessments: A White Paper.* Partnership for Assessment of Readiness for College and Careers. (Available on http://www.parcconline.org/technical-advisory-committee)

Brennan, R. L. (2013, April). *A Multivariate Generalizability Analysis of Portfolio Assessments in Dental Education.* (CASMA Technical Report No. 34). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
(Retrieved from https://education.uiowa.edu/casma)

Brennan, R. L. (December, 2016). *Nominal Weights in Multivariate Generalizability Theory.* (CASMA Research Report No. 50). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
(Retrieved from https://education.uiowa.edu/casma)

Brennan, R. L. (January, 2017). *Using G Theory to Examine Confounded Effects: The Problem of One.* (CASMA Research Report No. 51). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Retrieved from https://education.uiowa.edu/casma)

Brennan, R. L. (2020a). *Generalizability theory: Contributions, Challenges, and Future Prospects.* Division D Annual Linn Award Address presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Brennan, R. L. (January, 2020b). *Generalizability Theory References: The First*

Sixty Years. (CASMA Research Report No. 53). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Retrieved from https://education.uiowa.edu/casma)

Brennan, R. L. (in press). Generalizability theory. In B Clauser (Ed.)

Brennan, R. L., Yin, P., & Kane, M. T. (2003). Methodology for examining the reliability of group mean difference scores. *Journal of Educational Measurement, 40*(3), 207–230.

Briggs, D. C., & Alzen, J. L. (2019). Making inferences about teacher observation scores over time. *Educational and Psychological Measurement, 79*(4), 636–664.

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement, 44*(2), 131–155.

Broglio, S. P., Zhu, W., Sopiarz, K., & Park, Y. (2009). Generalizability theory analysis of balance error scoring system reliability in healthy young adults. *Journal of Athletic Training, 44*(5), 497–502.

Cardinet, J., Johnson. S., & Pini, G. (2010). *Applying generalizability theory using EduG.* New York, Routledge.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311–337.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783.

Chafouleas, S. M., Christ, T. J., & Riley-Tillman, T. C. (2009). Generalizability of scaling gradients on direct behavior ratings. *Educational and Psychological Measurement, 69*(1), 157–173.

Chang, L. (1997). Dependability of anchoring labels of Likert-type scales. *Educational and Psychological Measurement, 57*(5), 800–807.

Christ, T. J., Riley-Tillman, T. C., Chafouleas, S. M., & Boice, C. H. (2010). Direct Behavior Rating (DBR): Generalizability and dependability across raters and observations. *Educational and Psychological Measurement, 70*(5), 825–843.

Clark, S., & Rose, D. J. (2001). Evaluation of dynamic balance among community-dwelling older adult fallers: A generalizability study of the limits of stability test. *Archives of Physical Medicine and Rehabilitation, 82*(4), 468–474.

Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37*(3), 245–261.

Clauser, B. E., Harik, P., & Margolis, M. J. (2006). A multivariate generalizability analysis of data from a performance assessment of physicians clinical

skills. *Journal of Educational Measurement, 43*(3), 173–191.

Clauser, B. E., Kane, M. T., & Clauser, J. C. (in press). Examining the precision of cut scores within a generalizability theory framework: A closer look at the item effect. *Journal of Educational Mesurement*.

Clauser, J. C., Margolis, M. J., & Clauser, B. E. (2014). An examination of the replicability of Angoff standard setting results within a generalizability theory framework. *Journal of Educational Measurement, 51*(2), 127–140.

Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a perfomance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement*, *37*, 245–261.

Clauser, B. E., Swanson, D. B., & Harik, P. (2002). Multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39*(4), 269–290.

Cook, D. A., Beckman, T. J., Mandrekar, J. N., & Pankratz, V. S. (2010). Internal structure of mini-CEX scores for internal medicine residents: Factor analysis and generalizability. *Advances in Health Sciences Education, 15*(5), 633–645.

Corkum, P., Andreou, P., Schachar, R., Tannock, R., & Cunningham, C. (2007). The telephone interview probe: A novel measure of treatment response in children with attention deficit hyperactivity disorder. *Educational and Psychological Measurement, 67*(1), 169–185.

Cronbach, L. J.(2004). My current thoughts on coefficient alpha and successor procedures. (Editorial assistance provided by R. Shavelson.) *Educational and Psychological Measurement, 64*(3), 391–418.

Dagnone, J. D., Hall, A. K., Sebok-Syer, S., Klinger, D., Woolfrey, K., Davison, C., Ross, J., McNeil, G., & Moore, S. (2016). Competency-based simulation assessment of resuscitation skills in emergency medicine postgraduate trainees–a Canadian multi-centred study. *Canadian Medical Education Journal, 7*(1), e57.

de Oliveira Filho, G. R., Vieira, J. E., & Schonhorst, L. (2005). Psychometric properties of the Dundee Ready Educational Environment Measure (DREEM) applied to medical residents. *Medical Teacher, 27*(4), 343–347.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*(5), 783–801.

Fan, X., & Chen, M. (2000). Published studies of interrater reliability often overestimate reliability: Computing the correct coefficient. *Educational and Psychological Measurement, 60*(4), 532–542.

Gadbury-Amyot, C. C., McCracken, M. S., Woldt, J. L., & Brennan, R. (2012). Implementation of portfolio assessment of student competence in two dental school populations. *Journal of Dental Education, 76*(12), 1559–1571.

Gadbury-Amyot, C. C., McCracken, M. S., Woldt, J. L., & Brennan, R. L. (2014). Validity and reliability of portfolio assessment of student competence in two dental school populations: A four-year study. *Journal of Dental Education, 78*(5), 657–667.

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, *14*, 191–203.

Gao, X, Brennan, R. L., & Guo, F. (2015, August). *Modeling measurement facets and assessing generalizability in a large-scale writing assessment.* GMAC Research Reports, RR-15-01. Graduate Management Admission Council, Reston, Virginia.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all?. *Language Testing, 26*(4), 507–531.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, 15(2), 100–117.

Gugiu, M. R., Gugiu, P. C., & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of grades assigned to undergraduate research papers. *Journal of MultiDisciplinary Evaluation, 8*(19), 26–40.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.),*Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

Han, C. (2016). Investigating score dependability in English/Chinese interpreter certification performance testing: A generalizability theory approach. *Language Assessment Quarterly, 13*(3), 186–201.

Han, T., & Ege, . (2013). Using generalizability theory to examine classroom instructors analytic evaluation of EFL writing. *International Journal of Education, 5*(3), 20–35.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43–58.

Harrison, G. M. (2015). Non-numeric intrajudge consistency feedback in an Angoff procedure. *Journal of Educational Measurement, 52*(4), 399–418.

Huang, J. (2008). How accurate are ESL students holistic writing scores on large-scale assessments?–A generalizability theory approach. *Assessing Writing*, 13(3), 201–218.

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2*(4), 423–443.

Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing cutoffscores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement, 59*(6), 885–897.

Ishikawa, S., Kang, M., Bjornson, K. F., & Song, K. (2013). Reliably measuring

ambulatoryactivity levels of children and adolescents with cerebral palsy. *Archives of Physical Medicine and Rehabilitation, 94*(1), 132–137.

Jiang, Z., & Raymond, M. (2018). The use of multivariate generalizability theory to evaluate the quality of subscores. *Applied Psychological Measurement, 42*(8), 595–612.

Kane, M. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement, 39*(2), 165–181.

Kane, M. T. (2011). The errors in our ways. *Journal of Educational Measurement, 48* 12–30.

Kassab, S. E., Fida, M., Radwan, A., Hassan, A. B., Abu-Hijleh, M., & OConnor, B. P. (2016). Generalizability theory analyses of concept mapping assessment scores in a problem-based medical curriculum. *Medical Education, 50*(7), 730–737.

Kersting, N. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers knowledge of teaching mathematics. *Educational and Psychological Measurement, 68*(5), 845–861.

Kim, S. Y., Lee, W., & Brennan, R. L. (December, 2016). Reliability of mixed-format composite scores involving raters: A Multivariate generalizability theory approach. In M. J. Kolen & W. Lee (Eds.) *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating* (Volume 4). (CASMA Monograph 2.4). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Retrieved from https://education.uiowa.edu/casma)

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Kolen, M. J. & Harris, D. J. (April, 1987). *A multivariate test theory model based on item response theory and generalizability theory.* A paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics, 14*(2), 1–23.

Kreiter, C. D., Solow, C., Brennan, R. L., Yin, P., Ferguson, K., & Huebner, K. (2006). Examining the influence of using same versus different questions on the reliability of the medical school pre-admission interview. *Teaching and Learning in Medicine, 18*(1), 4–8.

Kreiter, C. D., Yin, P., Solow, C., & Brennan, R. L. (2004). Investigating the reliability of the medical school admissions interview. *Advances in Health Sciences Education, 9*, 147–159.

Lagha, R. A. R., Boscardin, C. K., May, W., & Fung, C. C. (2012). A comparison of two standard–setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine,*

87(8), 1077–1082.

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.

Lee, G. (2000a). A comparison of methods of estimating conditional standard errors of measurement for testlet-based test scores using simulation techniques. *Journal of Educational Measurement, 37*(2), 91-112.

Lee, G. (2000b). Estimating conditional standard errors of measurement for tests composed of testlets. *Applied Measurement in Education, 13*, 161-180.

Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. *Journal of Educational Measurement, 39*(2), 149-164.

Lee, G., & Fitzpatrick, A. R. (2003). The effects of a student sampling plan on estimates of the standard errors for student passing rates. *Journal of Educational Measurement, 40*(1), 17-28.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 120.

Lee, Y. W., Gentile, C., & Kantor, R. (2008). Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater. *ETS Research Report Series*, 2008(1), 1-71.

Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes. *ETS Research Report Series*, 2005(1), 1-76.

Lee, G., & Lewis, D. M. (2008). A generalizability theory approach to standard error estimates for bookmark standard settings. *Educational and Psychological Measurement, 68*(4), 603-620.

Li, D., & Brennan, R. L. (2007, August). A multi-group generalizability analysis of a large-scale reading comprehension test. (CASMA Research Report No. 25). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
(Retrieved from https://education.uiowa.edu/casma)

Li, M. N. F., & Lautenschlager, G. (1997). Generalizability theory applied to categorical data. *Educational and Psychological Measurement, 57*(5), 813-822.

Lin, C. K. (2017). Working with sparse data in rated language tests: Generalizability theory applications. *Language Testing, 34*(2), 271-289.

Lin, C. K., & Zhang, J. (2018). Detecting nonadditivity in single-facet generalizability theory applications: Tukeys test. *Journal of Educational Measurement, 55*(1), 78-89.

MacIntyre, N. J., Bennett, L., Bonnyman, A. M., & Stratford, P. W. (2011). Optimizing reliability of digital inclinometer and flexicurve ruler measures of spine curvatures in postmenopausal women with osteoporosis of the spine: An illustration of the use of generalizability theory. *ISRN rheumatology*, 571698.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement, 74*(3), 400-422.

Moses, T., & Kim, S. (2015). Methods for evaluating composite reliability, classification consistency, and classification accuracy for mixed–format licensure tests. *Applied Psychological Measurement, 39*(4), 314-329.

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi–trait scales: A generalizability analysis. *Assessing Writing, 38*, 21–36.

Oosterveld, P., & Cate, O. T. (2004). Generalizability of a study sample assessment procedure for entrance selection for medical school. *Medical Teacher, 26*(7), 635–639.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large–scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27*(4), 341–384.

Powers, S., & Brennan, R. L. (2009, September). Multivariate generalizability analyses of mixed-format exams. (CASMA Research Report No. 29). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
(Retrieved from https://education.uiowa.edu/casma)

Raymond, M. R., Harik, P., & Clauser, B. E. (2011). The impact of statistically adjusting for rater effects on conditional standard errors of performance ratings. *Applied Psychological Measurement, 35*(3), 235–246.

Raymond, M. R., & Jiang, Z. (2020). Indices of subscore utility for individuals and subgroups based on multivariate generalizability theory. *Educational and Psychological Measurement, 80*(1), 67–90.

Raymond, M. R., Swygert, K. A., & Kahraman, N. (2012). Psychometric equivalence of ratings for repeat examinees on a performance assessment for physician licensure. *Journal of Educational Measurement, 49*(4), 339–361.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*(1), 1–30.

Shavelson, R. J., & Webb, N. (2019). Generalizability theory and its contribution to the discussion of the generalizability of research findings. In K. Ercikan & W. Roth (Ed.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 13–32). New York:

Routledge.

Shin, S. Y., & Ewert, D. (2015). What accounts for integrated reading–to–write task scores?. *Language Testing, 32*(2), 259–281.

Shin, Y., & Raudenbush, S. W. (2012). Confidence bounds and power for the reliability of observational measures on the quality of a social setting. *Psychometrika, 77*(3), 543–560.

Smith, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement, 64*(4), 617–639.

Solano–Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13–22.

Strube, M. J. (2000). Reliability and generalizability theory. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23–66). Washington, DC: American Psychological Association.

Srinivasan, M., McElvany, M., Shay, J. M., Shavelson, R. J., & West, D. C. (2008). Measuring knowledge structure: Reliability of concept mapping assessment in medical education. *Academic Medicine, 83*(12), 1196–1203.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many–facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239–261.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E., Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*(3), 492–506.

Tavakol, M., & Brennan, R. L. (2013). Medical education assessment: A brief overview of concepts in generalizability theory. *International Journal of Medical Education, 4*, 221–222.

Tobar, D. A., Stegner, A. J., & Kane, M. T. (1999). The use of generalizability theory in examining the dependability of scores on the profile of mood states. *Measurement in Physical Education and Exercise Science, 3*(3), 141–156.

Tong, Y., & Brennan, R. L. (2004). *Bootstrap procedures for estimating standard errors of estimated variance components for two-facet designs.* (CASMA Research Report No. 5). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Retrieved from https://education.uiowa.edu/casma)

Tong, Y., & Brennan, R. L. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement, 67*(5), 804–817.

Trejo-Meja, J. A., Snchez-Mendiola, M., Mndez-Ramrez, I., & Martnez-Gonzlez, A. (2016). Reliability analysis of the objective structured clinical examination using generalizability theory. *Medical Education Online, 21*(1), 31650.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Using G-theory to enhance evidence of reliability and validity for common uses of the Paulhus Deception Scales. *Assessment, 25*(1), 69–83.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Using generalizability theory to disattenuate correlation coefficients for multiple sources of measurement error. *Multivariate Behavioral Research, 53*(4), 481–501.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *Journal of Personality Assessment, 100*(1), 53–67.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods, 23*(1), 1–26.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods, 24*(2), 153–178.

Vispoel, W. P., & Tao, S. (2013). A generalizability analysis of score consistency for the Balanced Inventory of Desirable Responding. Psychological Assessment, 25(1), 94–104.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement, 40*(3), 231–253.

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education 13*, 277–301.

Welk, G. J., Schaben, J. A., & Morrow Jr, J. R. (2004). Reliability of accelerometry-based activity monitors: A generalizability study. *Medicine & Science in Sports & Exercise, 36*(9), 1637–1645.

Wickel, E. E., & Welk, G. J. (2010). Applying generalizability theory to estimate habitual activity levels. *Medicine & Science in Sports & Exercise, 42*(8), 1528–1534.

Wiley, E. W. (2000). *Bootstrap strategies for variance component estimation: theoretical and empirical results.* Unpublished doctoral dissertation, Stanford.

Worster, A., Sardo, A., Eva, K., Fernandes, C. M., & Upadhye, S. (2007). Triage tool inter-rater reliability: A comparison of live versus paper case scenarios. *Journal of Emergency Nursing, 33*(4), 319–323.

Wu, Y. F., & Tzou, H. (2015). A multivariate generalizability theory approach to standard setting. *Applied Psychological Measurement, 39*(7), 507–524.

Yin, P. (2005). A multivariate generalizability analysis of the Multistate Bar

Examination. *Educational and Psychological Measurement, 65*(4), 668–686.

Yin, P., & Sconing, J. (2008). Estimating standard errors of cut scores for item rating and mapmark procedures: A generalizability theory approach. *Educational and Psychological Measurement, 68*(1), 25–41.

Zhang, J., & Lin, C. K. (2016). Generalizability theory with one-facet nonadditive models. *Applied Psychological Measurement, 40*(6), 367–386.