**RELIABILITY: SELECTED REFERENCES**

**(Last updated 14 November 2020)**

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125-141.

Berk, (1984). Selecting the index of reliability (pp. 231-266). In R. A. Berk (Ed.) *A guide to criterion-referenced test construction*. Baltimore, MD: Johns Hopkins University.

Biler, A. (2019). The reliability and readability tools in L2 reading. In S. Papageorgiou & K. M. Bailey (Eds.), *Global perspectives on language assessment: Research, theory, and practice* (pp. 108-121). New York, NY: Routledge.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*, 105–121.

#### Brown, J. D. (2001). Can we use the Spearman-Brown prophecy formula to defend low reliability?. JALT Testing & Evaluation SIG Newsletter, 4(3), 7-11.

Brown, J. D. (2002). The Cronbach alpha reliability estimate. *JALT Testing & Evaluation SIG Newsletter*, *6*(1), 17-19.

Brown, J. D. (2007). Multiple views of L1 writing score reliability. *Second Language Studies, 25*(2), 1-31. Retrieved from <http://www.hawaii.edu/sls/uhwpesl/25(2)/BrownWritingGstudy.pdf>

Berk, (1984). Selecting the index of reliability (pp. 231-266). In R. A. Berk (Ed.) *A guide to criterion-referenced test construction*. Baltimore, MD: Johns Hopkins University.

Blomert, L., Kean, M. L., Koster, C., & Schokker, J. (1994). Amsterdam—Nijmegen everyday language test: construction, reliability and validity. *Aphasiology*, *8*(4), 381-407.

Buck, G. (1992). Translation as a language testing procedure: Does it work?. *Language Testing*, *9*(2), 123-148.

Chiedu, R. E., & Omenogor, H. D. (2014). The concept of reliability in language testing: Issues and solutions. *Journal of Resourcefulness and Distinction*, *8*(1), 1-9.

Cho, D. (1999). A study on ESL writing assessment: Intra-rater reliability of ESL compositions. *Melbourne Papers in Language Testing*, *8*(1), 1-24.

Cole, K. N., Mills, P. E., & Dale, P. S. (1989). Examination of test-retest and split-half reliability for measures derived from language samples of young handicapped children. *Language, Speech, and Hearing Services in Schools*, *20*(3), 259-268.

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.

Davis, K. A. (1992). Validity and reliability in qualitative research on second language acquisition and teaching: Another researcher comments…. *TESOL Quarterly*, *26*(3), 605-608.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88-115.

Fouly, K. A., & Cziko, G. A. (1985). Determining the reliability, validity, and scalability of the graduated dictation test. *Language learning*, *35*(4), 555-566.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions?. *Child Language Teaching and Therapy*, *26*(1), 77-92.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing*, *14*(2), 113-139.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System*, *28*(1), 31-53.

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1258-1262.

Geffen, G., & Caudrey, D. (1981). Reliability and validity of the dichotic monitoring test for language laterality. *Neuropsychologia*, *19*(3), 413-423.

Haertel, E. H. (2006). Reliability. In R. L. Linn (Ed.), *Educational measurement* (4th ed., pp. 65-110). New York, NY: American Council on Education & Praeger.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication, 16*, 354-367.

Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly*, *20*(3), 559-562.

Hout, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*, 201-213.

Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation approaches in marketing: A generalizability theory framework for quantitative data. *Journal of Marketing Research*, *27,* 185-195.

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication, 18*, 229-249.

Kang, O., & Rubin, D. L. (2012). Intra-rater reliability of oral proficiency ratings. *International Journal of Educational and Psychological Assessment, 12*(1), 43-61.

Kember, D., & Leung, D. Y. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education*, *33*(4), 341-353.

Krzanowski, W. J., & Woods, A. J. (1984). Statistical aspects of reliability in language testing. *Language Testing*, *1*(1), 1-20.

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.

Liao, Y. F. (2004). Issues of validity and reliability in second language performance assessment. *online] Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *4*(2), 1-4.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587-604.

McWilliam, R. A., & Ware, W. B. (1994). The reliability of observations of young children's engagement: An application of generalizability theory. *Journal of Early Intervention*, *18*(1), 34-47.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing*, *20*(1), 26-56.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating argumentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing, 7*, 143-164.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538-553.

Polio, C., & Yoon, H. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics, 28*(1), 165**-**188.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18-39.

Roebroeck, M. E., Harlaar, J., & Lankhorst, G. J. (1993). The application of generalizability theory to reliability assessment: An illustration using isometric force measurements. *Physical Therapy*, *73*(6), 386-395.

Rybowiak, V., Garst, H., Frese, M., & Batinic, B. (1999). Error orientation questionnaire (EOQ): Reliability, validity, and different language equivalence. *Journal of Organizational Behavior*, 527-547.

Shale, D, (2004). Essay reliability: Form and meaning. In W. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76-96). New York, NY: The Modern Language Association of America.

Shaw, S. (2002). The effect of training and standardization on rater judgement and inter-rater reliability. *Research Notes, 9*, 13–17. Retrieved from <http://www.cambridgeesol.org/rs> \_notes/rs\_nts8.pdf

Shohamy, E. (1983). Rater reliability of the oral interview speaking test. *Foreign Language Annals*, *16*(3), 219-222.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27-33.

Subkoviak, M.J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*, 47-55.

Surface, E. A., & Dierdorff, E. C. (2003). Reliability and the ACTFL Oral Proficiency Interview: Reporting indices of interrater consistency and agreement for 19 languages. *Foreign Language Annals*, *36*(4), 507-519.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E., Reed, M., ... & 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.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology, 7*(6), 456–461. doi: 10.1037/h0074469

Thompson, I. (1995). A study of interrater reliability of the ACTFL oral proficiency interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals*, *28*(3), 407-422.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, *61*(1), 295-304.

Webb, N. M., Shavelson, R. J., &Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics*, *26*, 81-124.

Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological measurement*, *34*(1), 25-33.

Wetherby, A. M., Allen, L., Cleary, J., Kublin, K., & Goldstein, H. (2002). Validity and reliability of the communication and symbolic behavior scales developmental profile with very young children. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1202-1218.

Whittington, D. (1999). Making room for values and fairness: Teaching reliability and validity in the classroom context. *Educational Measurement: Issues and Practice*, *18*(1), 14-22.

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System*, *35*(2), 182-191.