**VALIDITY AND VALIDATION IN ASSESSMENT AND RESEARCH:**

**SELECTED REFERENCES**

**(Last updated 27 March 2023)**

Abella, R., Urrutia, J., & Shneyderman, A. (2005). An examination of the validity of English-language achievement test scores in an English language learner population. *Bilingual Research Journal*, *29*(1), 127-144.

Aita, S. L., Beach, J. D., Taylor, S. E., Borgogna, N. C., Harrell, M. N., & Hill, B. D. (2018). Executive, language, or both? An examination of the construct validity of verbal fluency measures. *Applied Neuropsychology: Adult, 26*(5), 441-451. <https://doi.org/10.1080/23279095.2018.1439830>.

Alderson, J. C. (1988). New procedures for validating proficiency tests of ESP? Theory and practice.*Language Testing, 5*(2), 220-232.

Allison, D., & Cheung, E. (1991). ‘Good’ and ‘poor’ writing and writers: Studying individual performance as a part of placement test validation. *Hong Kong Papers in Linguistics and Language Teaching, 14,* 1-14.

Al Noor, H. (2020). A probe into the different aspects of ‘validity’ and ‘reliability’ of IELTS writing test. *International Journal of English Literature and Social Sciences, 5*(4), 968-972.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. D. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41-66.

Arkoudis, S., & O’Loughlin, K. (2004). Tensions between validity and outcomes: Teachers’ assessment of written work of recently arrived immigrant ESL students. *Language Testing, 20*, 284-304.

Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Cambridge Scholars Publishing.

Ayers, J. B., & Peters, R. M. (1977). Predictive validity of the test of English as a foreign language for Asian graduate students in engineering, chemistry, or mathematics. *Educational and Psychological Measurement*, *37*(2), 461-463.

Bachman, L. F. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Bachman, L. F. (1988). Problems in examining the validity of the oral proficiency interview. *Studies in Second Language Acquisition, 10*, 149-164.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1981). The construct of validation of the FSI oral interview. *Language Learning, 31*, 167-186.

Bachman, L. F., & Palmer, A. S. (1981). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence,* (pp. 149-165). TESOL.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449-465.

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*(1), 14-29.

Baik, S. H., Fox, R. S., Mills, S. D., Roesch, S. C., Sadler, G. R., Klonoff, E. A., & Malcarne, V. L. (2019). Reliability and validity of the Perceived Stress Scale-10 in Hispanic Americans with English or Spanish language preference. *Journal of Health Psychology*, *24*(5), 628-639.

Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), *Language testing and assessment. Encyclopedia of Language and Education* (Vol. 7, pp. 275-287). Kluwer.

Bateman, H. (2010). A study of the context and cognitive validity of a BEC vantage test of writing. *Cambridge ESOL Research Notes*, *42*, 40.

Beaudrie, S., Amezcua, A., & Loza, S. (2019). Critical language awareness for the heritage context: Development and validation of a measurement questionnaire. *Language Testing*, *36*(4), 573-594.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test.*Language Testing, 27*, 101-118.

Behizadeh, N., & Engelhard Jr., G. (2014). Development and validation of a scale to measure perceived authenticity in writing. *Assessing Writing, 21*, 18-36. Doi: [10.1016/j.asw.2014.02.001](https://doi.org/10.1016/j.asw.2014.02.001)

Bejar, I. I. (2011). A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 319-341.

Bennett, R. E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring*. Lawrence Erlbaum.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It’s not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Benson, J., Moulin-Joulin, M., Schwarzer, C., Seipp, B. & El-Zahhar, N. (1992). Cross validation of a revised test anxiety scale using multi-national sample. In K. A. Hagtver & T. B. Johnson (Eds.), *Advances in test anxiety research* (pp. 62-83). Swette & Zeitlinger.

Bers, T. H., & Smith, K. E. (1990). Assessing assessment programs: The theory and practice of examining reliability and validity of a writing placement test. *Community College Review*, *18*(3), 17-27.

Bhuyan Boruah, P. (2022). Visibility as validation: A case study of culturally responsive materials development for TESOL. *The CATESOL Journal, 33*(1). <http://www.catesoljournal.org/wp-content/uploads/2022/10/CJ33-1_Boruah.pdf>

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

Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, *70*(1), 11-47.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.

Bostancıoğlu, A., & Handley, Z. (2018). Developing and validating a questionnaire for evaluating the EFL ‘Total PACKage’: Technological Pedagogical Content Knowledge (TPACK) for English as a Foreign Language (EFL). *Computer Assisted Language Learning*, *31*(5-6), 572-598.

Breeze, R., & Miller, P. (2012) Predictive validity of the IELTS listening test as an indicator of student coping ability in English-medium undergraduate courses in Spain. In L. Taylor & C. Weir (Eds.), *Studies in Language Testing 34: Research in reading and listening assessment* (pp. 487-518). Cambridge University Press.

Brennan, R. L. (Ed.). (2006). *Educational measurement, 4th Ed.* American Council on Education.

Bridges, G. (2010). Demonstrating cognitive validity of IELTS academic writing task 1. *Cambridge ESOL Research Notes*, *42,* 24-33.

Brown, A. N., Dewey, D. P. & Cox, T. L. (2014). Assessing the validity of can-do statements in retrospective (then-now) self-assessment. *Foreign Language Annals, 47*(2), 261-285.

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. (2000). What is construct validity? *JALT Testing and Evaluation SIG Newsletter* 4(2), 7-10.

Brown, J. D. (2005). Language test validity. *Testing in language programs: A comprehensive guide to English language assessment* (pp. 220-251). McGraw-Hill.

Brown, J. D., Cunha, M. I. A., & Frota, S. (2001). The development and validation of a Portuguese version of the motivated strategies for learning questionnaire. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 257-280). University of Hawaii Press.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. (pp. 45–78). Hampton Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage.

Castro, S., & Lima, C. (2010). Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody.*Behavior Research Methods, 42*(1), 74-81. <http://ulib.iupui.edu/cgi-bin/proxy.pl?url=/docview/204304709?accountid=7398>

Chapelle, C. (1998) Construct definition and validity inquiry in SLA research. In L. Bachman & A. Cohen (Eds.), *Second language acquisition and language testing interfaces* (pp. 32-70). Cambridge University Press.

Chappelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254-272. doi:10.1017/S0267190599190135

Chapelle, C. (2011). Validation in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 717-730). Routledge.

Chapelle, C.A. (2012).  Validity argument for language assessment:  The framework is simple…  *Language Testing 29*(1), 19-27.

Chapelle, C. A. (2012). Conceptions of validity. In G. Flucher, & F. Davidson (Eds.), *Routledge Handbook of Language Testing,* (pp. 21-33).Routledge.

Chapelle, C. A., Enright, M. & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language™.* Routledge.

Chapelle, C. A. (2020). *Argument-based validation in testing and assessment*. Sage Publications.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). Routledge.

Chapelle, C. A., Enright, M. E., & Jamieson, J.  (2010).  Does an argument-based approach to validity make a difference?  *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, pp. 1079-1097). Wiley-Blackwell.

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity argument in language testing: Case studies of validation research.* Cambridge University Press.

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for test validation and test use. Educational Assessment, 16(2), 104-122.

Chapelle, C. A., & Lee, H. W. (2021). Conceptions of validity. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 17-31). Routledge.

Chen, Y., Zhong, R., Hu, H., Zhang, H., Yang, Y., Wu, D., & Lee, W. (2021, May). One engine to fuzz’em all: Generic language processor testing with semantic validation. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 642-658). IEEE.

Cheng, L., & Sun, Y. (2015). Interpreting the impact of the Ontario Secondary School Literacy Test on second language students within argument-based validation framework. Language Assessment Quarterly, 12, 50–66. doi: 10.1080/15434303.2014.981334

Cheng, Y. S. (2004). A measure of second language writing anxiety: Scale development and preliminary validation. *Journal of Second Language Writing,* *13*(4), 313-335. Doi: [10.1016/j.jslw.2004.07.001](https://doi.org/10.1016/j.jslw.2004.07.001)

Chudowsky, N., & Behuniak, P. (1997). *Establishing consequential validity for large-scale performance assessments*. National Council of Measurement in Education.

Cicourel, A. (2007). A personal, retrospective view of ecological validity. *Text & Talk*, 27, 735–752.

Clark, J. L. D. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency. *Language Testing, 5*, 187-205.

Clifford, A. (2004). Testing the test: A look at construct validity in interpreter certification/ Mettre l’examen à l’épreuve : La validité de construits et l’agrément d’interprètes. *The Critical Link – A Journal Dedicated to Interpreting in the Social, Health Care and Legal Sectors*, 11(2), 21-26. Critical Link Canada. http://www.criticallink.org/English/index2.htm

Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting 7*(1), 97-131.

Coady, M., Miller, M. D., Jing, Z., Qlszewska, A., De Jong, E., Yilmaz, T., Heffington, D., Lopez, M., & Ankeny, R. (2020). Can English learner teacher effectiveness be observed? Validation of an EL-modified framework for teaching. *TESOL Quarterly, 54*(1), 173-200.

Collie, R. J., Martin, A. J., & Curwood, J. S. (2016). Multidimensional motivation and engagement for writing: Construct validation with a sample of boys. *Educational Psychology, 36*(4). 771-791. Doi: [10.1080/01443410.2015.1093607](https://doi.org/10.1080/01443410.2015.1093607)

Cox, T. L. & Clifford, R. (2014). Empirical validation of listening proficiency guidelines. *Foreign Language Annals, 47*(3), 379-403.

Cox, T., & Malone, M. (2018). A validity argument to support the ACTFL assessment of performance toward proficiency in languages (AAPPL). *Foreign Language Annals, 51*(3), 548-574.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3-17). Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validity after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302

Cumming, A. (1996). Introduction: The concept of validation in language testing. In A. Cumming & R. Berwick (Eds.), *Validation in Language Testing* (pp. 1-14). Multilingual Matters.

Cumming, A., & Berwick, R. (Eds.). (1996). *Validation in language testing.* Multilingual Matters Ltd.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 72-93). Multilingual Matters.

Cushing Weigle, S., & Lynch, B. (1996). Hypothesis testing in construct validation. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 58-71). Multilingual Matters.

Dahllöf, U. S. (1971). *Ability grouping, content validity and curriculum process analysis.* Teachers College Press.

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*, 11-22.

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General*, *130*(2), 208.

Darbes, T. (2016). The implications of test taker perceptions for test validity in community college settings. In M. A. Christison & N. Saville (Eds.), *Advancing the field of language assessment: Papers from TIRF doctoral dissertation grantees. Studies in Language Testing 46* (pp. 193-211). Cambridge University Press.

Davari, H., Karami, H., Nourzadeh, S., & Iranmehr, A. (2022). Examining the validity of the Achievement Emotions Questionnaire for measuring more emotions in the foreign language classroom. *Journal of Multilingual and Multicultural Development*, *43*(8), 701-714.

Davies, A. (1996). The role of the segmental dictionary in professional validation: Constructing a dictionary of language testing. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 222-235). Multilingual Matters.

Davies, A., & Elder, C. (2011). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 705-813). Routledge.

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

Delgado, C., Araneda, A., & Behrens, M. I. (2019). Validation of the Spanish-language version of the Montreal Cognitive Assessment test in adults older than 60 years. *Neurología (English Edition)*, *34*(6), 376-385.

Desvousages, W. H., Johnson, F. R., Dunford, R. W., Boyle, K. J., Hudson, S. P., & Wilson, K. N. (1993). Measuring natural resource damages with contingent valuation: Tests of validity and reliability. In J. Hausman (Ed.), *Contingent valuation: A critical assessment* (pp. 91-164). North-Holland Press.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. A. Chapelle, & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 9-25). John Benjamins.

Dobrić, N. (2018). Reliability, validity, and writing assessment: A timeline. *ELOPE: English Language Overseas Perspectives and Enquiries*, *15*(2), 9-24.

Dooey, P., & Oliver, R. (2002). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. *Prospect, 17*(1), 36-54.

DuFon, M. A. (2002). Video recording in ethnographic SLA research: Some issues of validity in data collection. *Language Learning & Technology*, *6*(1), 40–59.

Dunlea, J., Spiby, R., Nguyen, T., Quynh, N., Nguyen, T., Huu, M., Nguyen, T., Yen, Q., & Thai, H. L. (2018). *Aptis-VSTEP comparability study: Investigating the usage of two EFL tests in the context of higher education in Vietnam* (British Council Validations Series VS/2018/001). British Council.

Duran, R. P. (1988). Validity and language skills assessment: Non-English background students. *Test validity*, 105-127.

Ebrahimi, N. (2015). Validation and application of the Constructivist Learning Environment Survey in English language teacher education classrooms in Iran. *Learning Environments Research*, *18*(1), 69-93.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290-325.

Elder, C., & Wigglesworth, G. (2006). An investigation of the effectiveness and validity of planning time in Part 2 of the IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELT Research reports Volume 6* (pp. 13-40). IELTS Australia and the British Council.

Elliott, M. & Wilson, J. (2011). Context validity. In L. Taylor (Ed.), *Studies in language testing, 30:* *Examining speaking: Research and practice in assessing second language speaking* (pp. 152-241). UCLES/Cambridge University Press.

Enright, M. K., Bridgeman, B., Eignor, D., Lee, Y. W., & Powers, D. E. (2008). Prototyping measures of listening, reading, speaking, and writing. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 145–186). Routledge.

Erickan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments*. Routledge.

Evard, B. L., & Sabers, D. L. (1979). Speech and language testing with distinct ethnic-racial groups: A survey of procedures for improving validity. *Journal of Speech and Hearing Disorders*, *44*(3), 271-281.

Fan, J. (2016). The construct and predictive validity of a self-assessment scale. *Papers in Language Testing and Assessment*, *5*(2), 69-100.

Fang, X., & Yang, H. (2017). Validity and validation of language proficiency scales. *Journal of Foreign Languages*, *40*(4), 2-14.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*(3), 274-291.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Studies in language testing, 30:* *Examining speaking: Research and practice in assessing second language speaking* (pp. 65–111). UCLES/Cambridge University Press.

Fiske, S. I., Haddeland, A. L., Skipar, I., Bootsma, J. N., Geytenbeek, J. J., & Stadskleiv, K. (2020). Assessing language comprehension in motor impaired children needing AAC: validity and reliability of the Norwegian version of the receptive language test C-BiLLT. *Augmentative and Alternative Communication*, *36*(2), 95-106.

Fitzpatrick, T., & Clenton, J. (2010). The challenge of validation: Assessing the performance of a test of productive vocabulary.*Language Testing, 27*, 537-554.

Forbes, C. T., Zangori, L., & Schwarz, C. (2015). Empirical validation of integrated learning performances for hydrologic phenomena: 3rd-grade students’ model-driven explanation-construction. *Journal of Research in Science Teaching, 52*(7), 895-921.

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

Frederiksen, N. (1986). Construct validity and construct similarity: Methods for use in test development and test validation. *Multivariate Behavioral Research*, *21*(1), 3-28.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*(1), 2-32.

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.

Fulcher, G. (1999). Assessment in English for academic purposes: Putting content validity in its place. *Applied Linguistics, 20*, 221-236.

Galante, A. (2015). Intercultural communicative competence in English language teaching: Towards validation of student identity. *Brazilian English Language Teaching Journal, 6*(1), 29-39.

Garver, M. S., & Mentzer, J. T. (1999). Logistics research methods: employing structural equation modeling to test for construct validity. *Journal of Business Logistics*, *20*(1), 33.

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

Gellert, A., & Carsten, E. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31*(1), 16-28.

Geranpayeh, A. (2011). Scoring validity. In L. Taylor (Ed.), *Studies in language testing, 30:* *Examining speaking: Research and practice in assessing second language speaking* (pp. 242-272). UCLES/Cambridge University Press.

Goh, S. K., Tham, E. K., Magiati, I., Sim, L., Sanmugam, S., Qiu, A., ... & Rifkin-Graboi, A. (2017). Analysis of item-level bias in the Bayley-III language subscales: The validity and utility of standardized language assessment in a multilingual setting. *Journal of Speech, Language, and Hearing Research*, *60*(9), 2663-2671.

Greene, J. C. (2011). The construct(ion) of validity as argument. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing validity in outcome evaluation: Theory and practice, new directions for evaluation* (pp. 81-92). Jossey-Bass.

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. *Language Testing, 3*, 159-185.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23,* 17-27.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. *Adapting Educational and Psychological Tests for Cross-cultural Assessment*, *1*, 3-38.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment*, *11*(3), 147.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, *1*(1), 1-13.

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers’ perspectives on a global test of English: Questions of fairness, justice and validity. *Language Testing in Asia*, *9*, 1-20.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language testing*, *14*(3), 295-303.

Hamp-Lyons, L., & Lynch, B. K. (1998). Perspectives on validity: A historical analysis of language testing conference abstracts. In A. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th language testing research colloquium* (pp. 253-276). Lawrence Erlbaum.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests for individuals with disabilities within a validity framework. *System, 33*(1), 107-133.

Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*, 228-250.

Hashemi, A., & Daneshfar, S. (2018). A review of the IELTS test: Focus on validity, reliability, and washback. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, *3*(1), 39-52.

Hawkey, R. (2011). Consequential validity. In L. Taylor (Ed.), *Studies in language testing, 30:* *Examining speaking: Research and practice in assessing second language speaking* (pp. 273-302). UCLES/Cambridge University Press.

He, L. Z., &Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing 23*, 370-401.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language Learning, 33,* 315-332.

Henning, G. (1992). The ACTFL oral proficiency interview: Validity evidence. *SYSTEM, 20*, 365-372.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1-11.

Hill, K. (1997). The role of questionnaire feedback in the validation of the oral interaction module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery* (pp. 147-174). NCELTR.

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Language Assessment Quarterly*, *10*(4), 398-422.

Hoeve, K. B. (2022). A validity framework for accountability: Educational measurement and language testing. *Language Testing in Asia*, *12*(1), 1-14.

Holliday, A. (2004). Issues in validity in progressive paradigms of qualitative research. *TESOL Quarterly, 38*(4), 731-734.

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

Huang, H. T. D., Hung, S. T. A., Chao, H. Y., Chen, J. H., Lin, T. P., & Shih, C. L. (2022). Developing and validating a computerized adaptive testing system for measuring the English proficiency of Taiwanese EFL university students. *Language Assessment Quarterly*, *19*(2), 162-188.

Huong, T. T. T. (2001). The predictive validity of the international English Language Testing System (IELTS) test. *Post-Script*, *2*(1), 66-96.

Hwang, K. Y., & Dizney, H. F. (1970). Predictive validity of the test of English as a foreign language for Chinese graduate students at an American university. *Educational and Psychological Measurement*, *30*(2), 475-477.

Im, G.-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia, 9*(14). <https://doi.org/10.1186/s40468-019-0089-4>

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review, 64*(4), 555-580.

Jafarpur, A. (1996). Native speaker performance validity: In vain or for gain? *System, 24*(1), 83-95.

Jamalzadeh, M., Lotfi, A. R., & Rostami, M. (2021). Assessing the validity of an IAU General English Achievement Test through hybridizing differential item functioning and differential distractor functioning. *Language Testing in Asia*, *11*, 1-17.

Janssen, N., Roelofs, A., van den Berg, E., Eikelboom, W. S., Holleman, M. A., in de Braek, D. M., ... & Kessels, R. P. (2022). The diagnostic value of language screening in primary progressive aphasia: Validation and application of the Sydney Language Battery. *Journal of Speech, Language, and Hearing Research*, *65*(1), 200-214.

Janssen-van Dieten, A. M. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing*, *6*(1), 30-46.

Jenkins, J., & Leung, C. (2019). From mythical ‘standard’to standard reality: The need for alternatives to standardized English language tests. *Language Teaching*, *52*(1), 86-110.

Jin, Y. & Cheng, L. (2013). The effects of psychological factors on the validity of high-stakes test. *Modern Foreign Languages (Quarterly), 36*(1), 62-69.

Jing, X. (2019). The reliability and validity of language proficiency assessments for English language learners. *Frontier of Higher Education*, *1*(1), 36-42.

Johnson, D. M., & Saville-Troike, M. (1992). Validity and reliability in qualitative research on second language acquisition and teaching. *TESOL Quarterly, 26*, 602-605.

Jones, E. (2006). Accuplacer’s essay-scoring technology: When reliability does not equal validity. In P. F. Ericsson & R.H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 93-113). Utah State University Press

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practices, 21*(1), 31-41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective, 2(3)*, 135-170.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (pp. 17-64). American Council on Education and Praeger.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). National Council on Measurement in Education & American Council on Education.

Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing, 29*, 3-17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. doi:10.1111/jedm.12000

Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*(1), 115-122. doi:10.1111/jedm.12007

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

Kapborg, I., & Berterö, C. (2002). Using an interpreter in qualitative interviews: Does it threaten validity?. *Nursing Inquiry*, *9*(1), 52-56.

Kavanoz, S., & Yüksel, G. (2016). Developing and validating a self-efficacy scale for scholarly writing in English. *International Online Journal of Educational* *Sciences,* *8*(2), 71-82.

Khalifa, H. (2010). Construct validation of the reading module of an EAP proficiency test battery. *Cambridge ESOL Research Notes*, *42*, 8-14.

Khatib, M., & Nourzadeh, S. (2015). Development and validation of an instructional willingness to communicate questionnaire. *Journal of Multilingual and Multicultural Development*, *36*(3), 266-283.

Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *79*(3), 1168-1201.

King, R. B., & Ganotice Jr., F. A. (2013). Student motivation as hierarchical and multidimensional: Cross-cultural validation of personal investment theory in the Philippines. *Universitas Psychologica*, *12*(3), 1-32. <https://doi.org/10.11144/javeriana.upsy12-3.smhm>

Kim, Y. M., Yun, J. H., Lee, B. C., & Park, J. S. (2012). *Validating 2012 English reading and listening test items for College Scholastic Ability Test*. Korea Institute for Curriculum and Evaluation.

Klassen, R. M., Bong, M., Usher, E. L., Chong, W. H., Huan, V. S., Wong, I. Y., & Georgiou, T. (2009). Exploring the validity of a teachers’ self-efficacy scale in five countries. *Contemporary Educational Psychology*, *34*(1)*,* 67-76. doi:10.1016/j.cedpsych.2008.08.001

Klee, T., Carson, D. K., Gavin, W. J., Hall, L., Kent, A., & Reece, S. (1998). Concurrent and predictive validity of an early language screening program. *Journal of Speech, Language, and Hearing Research*, *41*(3), 627-641.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477-499.

Knoch, U., & Elder, C. (2010). Validity and fairness implications of varying time conditions on a diagnostic test of academic English writing proficiency. *System*, *38*(1), 63-74.

Kunnan, A. (1998). Approaches to validation in language assessment. In A. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th language testing research colloquium, Long Beach* (pp. 1-16). Lawrence Erlbaum.

Kunnan, A. (1998). Preface. In A. Kunnan (Ed.), *Validation in language assessment: Selected papers from the 17th language testing research colloquium, Long Beach* (pp. ix-x). Lawrence Erlbaum.

Kunnan, A. (Ed.). (1998). *Validation in language assessment: Selected papers from the 17th language testing research colloquium, Long Beach.* Lawrence Erlbaum.

Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (Vol. 9). Cambridge University Press.

Kutuk, G., Putwain, D. W., Kaye, L., & Garrett, B. (2020). Development and validation of a new multidimensional language class anxiety scale. *Journal of Psychoeducational Assessment*, *38*(5), 649-658.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(4), 451-475.

Latif, M. M. A. (2013). What do we mean by writing fluency and how can it be validly measured?. *Applied linguistics*, *34*(1), 99-105.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests.*  Cambridge University Press.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294-304.

Lee, B. Y., & Shin, S. K. (2020). Doable and practical: A validation study of classroom diagnostic tests. *Journal of Asia TEFL*, *17*(2), 363-378.

Lee, H. K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, *24*(3), 307-330.

Lee, Y. (2007). Validity and reliability analyses of the language test for school-age children. *Communication Sciences & Disorders*, *12*(4), 569-586.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, *40*(1), 211-234.

Li, C., Jiang, G., & Dewaele, J. M. (2018). Understanding Chinese high school students’ foreign language enjoyment: Validation of the Chinese version of the foreign language enjoyment scale. *System*, *76*, 183-196.

Li, X. (2018). Self-assessment as ‘assessment as learning’ in translator and interpreter education: Validity and washback. *The Interpreter and Translator Trainer*, *12*(1), 48-67.

Li, Z. (2015). Using an English self-assessment tool to validate an English placement test. *Papers in Language Testing and Assessment*, *4*(1), 59-96.

Liao, Y. F. (2022). Using the English GSAT for placement into EFL classes: Accuracy and validity concerns. *Language Testing in Asia*, *12*(1), 1-23.

Lim, G. (2013). Components of an elaborated approach to test validation. *Cambridge English Research Notes, 51*, 11-14.

Lim, G. S. (2012). Developing and validating a mark scheme for writing. *Research Notes, 49*, 6-1.

Lim, G.S. & Khalifa, H. (2011). Criterion-related validity. In L. Taylor (Ed.), *Studies in language testing, 30:* *Examining speaking: Research and practice in assessing second language speaking* (pp. 303-321). UCLES/Cambridge University Press.

Lin, C. K., & Zhang, J. (2013). Enhancing standard-based validity for ELL population: A perspective from correspondence between standards. *TESOL Quarterly, 47*(2), 399-410.

Lindquist, E. F. (Ed.). (1951). *Educational measurement.* American Council on Education.

Linn, R. L. (Ed.). (1989). *Educational measurement, 3rd Ed*. American Council on Education.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*, 28-30.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

Liu, T., Aryadoust, V., & Foo, S. (2022). Examining the factor structure and its replicability across multiple listening test forms: Validity evidence for the Michigan English Test. *Language Testing*, *39*(1), 142-171.

Liu, Z., Li, T., & Diao, H. (2020). Analysis on the reliability and validity of teachers' self-designed English listening test. *Journal of Language Teaching and Research*, *11*(5), 801-808.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489-515.

Long, M. H. (1997). Construct validity in SLA research. *The Modern Language Journal, 81,* 318-23.

Low, G. (1985). Validity and the problem of direct language proficiency tests. In J. C. Alderson (Ed.), *Lancaster papers in English language education:* *Evaluation* (pp. 151-168). Pergamon Press.

Lowe, P. (1981). Structure of the oral interview and content validity. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 71-80). TESOL.

Lyons, N. (Ed.). (1998). *With portfolio in hand: Validating the new teacher professionalism.* Teachers College Press.

MacSwan, J., Rolstad, K., & Glass, G. V. (2002). Do some school-age children have no language? Some problems of construct validity in the Pre-LAS Español. *Bilingual Research Journal*, *26*(2), 395-420.

Maerten-Rivera, J., Huggins-Manley, A. C., Adamson, K., Lee, O., & Llosa, L. (2015). Development and validation of a measure of elementary teachers’ science content knowledge in two multi-year teacher professional development intervention projects. *Journal of Research in Science Teaching,* *52*(3), 371-396.

Mahon, E. A. (2006). High-stakes testing and English language learners: Questions of validity. *Bilingual Research Journal*, *30*(2), 479-497.

Markham, K. M., Mintzes, J. J., & Gail Jones, M. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching*, *31*, 91-101.

Markus, K. A. (1998). Measurement and validity: Is completion of Samuel Messick’s synthesis possible? *Social Indicators Research, 45*(1/3), 7-34.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology, 77*, 413-440. Doi: [10.1348/000709906X118036](https://doi.org/10.1348/000709906X118036)

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31-51.

McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, justice & language assessment*. Oxford University Press.

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing*, *35*, 41-55.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 25*, 1012-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning of consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13- 103). Macmillan.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1996). Standards-based score interpretation: Establishing valid grounds for valid inferences. *Proceedings of the joint conference on standard setting for large scale assessments.* Sponsored by National Assessment Governing Board and The National Center for Education Statistics. Government Printing Office.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256.

Messick, S. (1996). Validity of performance assessment. In G. W. Philips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). National Center for Educational Statistics.

Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 3-20). Kluwer.

Mickan, P., Slater, S., & Gibson, C. (2000). Study of response validity of the IELTS Writing Subtest. In R. Tulloh (Ed.), *IELTS research reports,* *Vol. 3* (pp. 29-48). IELTS Australia.

Min, S., He, L., & Luo, L. (2018). Validation of listening descriptors of China’s Standards of English: An analysis of self-assessment data using polytomous IRT models. *Foreign Languages in China*, *15*(2), 72-81.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, *36*(8), 463-469.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13-25.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109-162.

Nicholson, S. J. (2015). Evaluating the TOEIC® in South Korea: Practicality, reliability and validity. *International Journal of Education*, *7*(1), 221-233.

Norris, J. M. (2008). *Validity evaluation in language assessment*. Peter Lang.

Obeidat, M. F. A., & Sheik, A. (2021). An evaluation of the national English language exam in Jordan for postgraduate studies: Validity and reliability. *International Journal of English Language Education, 9*(2), 96-113.

O'Dwyer, J., Kantarcıoğlu, E., & Thomas, C. (2018). An investigation of the predictive validity of the TOEFL iBT® test at an English-medium university in Turkey. *ETS Research Report Series, 2018*(1), 1-13.

O’Sullivan, B., & Weir, C. J. (2011). Language testing and validation**.** In B. O’Sullivan (Ed.), *Language testing: Theory & practice* (pp. 13-32). Palgrave.

Phakiti, A. (2008). Construct validation of Bachman and Palmer’s (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25*(2), 237–272.

Palmer, A. S., Groot, P. J. M., & Trosper, G. A. (Eds.). (1981). *The construct validation of tests of communicative competence.* TESOL.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, *12*(2), 153-177.

Pearlman, M. (2008). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 227- 258). Routledge.

Petric, B., & Czarl, B. (2003). Validating a writing strategy questionnaire. *System*, *31*, 187-215. doi:10.1016/S0346-251X(03)00020-4

Piazza, C., & Siebert, C. (2008). Development and validation of a writing dispositions scale for elementary and middle school students. The Journal of Educational Research, 101, 275-285. Doi: [10.3200/JOER.101.5.275-286](https://doi.org/10.3200/JOER.101.5.275-286)

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.

Pomerantz, A. (1990). Chautauqua: On the validity and generalizability of conversational analysis methods: Conversation analytic claims. *Communication Monographs, 57*, 231-235.

Powers, D. E. (2010). *Validity: What does it mean for the TOEIC tests?* (TOEIC Compendium Study TC-10-01). Educational Testing Service.

Powers, D. E., Mercadante, R., & Yan, F. (2013). Validating TOEIC Bridge scores against teacher ratings for vocational students in China. In D. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies: Volume II* (pp. 4.1-4.11). Educational Testing Service.

Powers, D. E., Schedl, M. A., Wilson-Leung, S. W., & Butler, F. A. (1999). Validating the Revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*(4), 399-425. doi:10.1177/026553229901600401

Powers, D. E., & Simpson, A. G. (2008). *Validating TOEIC Bridge scores against teacher and student ratings: A small-scale study* (ETS Research Memorandum RM-08-03). Educational Testing Service.

Powers, D. E., & Yan, F. (2013). TOEIC Bridge scores: Validity evidence from Korea and Japan. In D. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies: Volume II* (pp. 5.1-5.10). Educational Testing Service.

Rahmatollahi, M., & Mohamadi Zenouzagh, Z. (2021). Designing and validating an evaluation inventory for assessing teachers’ professional accountability. *Language Testing in Asia*, *11*(1), 1-21.

Reckase, M. (1998). Consequential validity from the test developer’s perspective. *Educational Measurement: Issues and Practice, 17*, 13-16.

Restrepo, M. A., & Silverman, S. W. (2001). Validity of the Spanish Preschool Language Scale-3 for use with bilingual children. *American Journal of Speech-Language Pathology*, *10*(4), 382-393.

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

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing, 23,* 229-256.

Rohmah, N. (2019). Validity and reliability study on teacher-made assessment for English mid-term examination. *Advances in Social Science, Education and Humanities Research,* 254, 107-110.

Rose, K. R. (1994). On the validity of discourse completion tests in non-Western contexts. *Applied Linguistics, 15*, 1-14.

Saeedi, M. (2016). Construct validity of multiple-choice cloze test and cloze-elide test in testing reading comprehension among Iranian EFL Learners. *International Research Journal of Applied Basic Science*, *10*(6), 632-637.

Saeedi, M., Tavakoli, M., Kazerooni, S. R., & Parvaresh, V. (2011). Do c-test and cloze procedure measure what they purport to be measuring? A case of criterion-related validity. *International Journal of Cognitive and Language Sciences*, *5*(2), 190-199.

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439-462.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *1*, 39-82.

Sakyi, A. (2000). Validation of holistic scoring for writing assessment: How raters evaluate ESL compositions. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129-152). Cambridge University Press.

Schmidgall, J. (2017). *Articulating and evaluating validity arguments for the TOEIC® tests* (Research Report No. RR-17-51). ETS. <https://doi.org/10.1002/ets2.12182>

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109-120.

Schrauf, R. W., Weintraub, S., & Navarro, E. (2006). Is a validation of the Word Accentuation Test (WAT) of premorbid intelligence necessary for use among older, Spanish-speaking immigrants in the United States?  Journal of the International Neuropsychological Society, 12, 391-399.

Shaver, J. (1990). Reliability and validity of measures of attitudes toward writing and toward writing with the computer. *Written Communication, 7*(3), 37-392. Doi: [10.1177/0741088390007003004](https://doi.org/10.1177%2F0741088390007003004)

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405-450.

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.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shohamy, E. (1994). The role of language tests in the construction and validation of second-language acquisition theories. In E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 133-142). Lawrence Erlbaum.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11,* 99-123.

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *Modern Language Journal, 95*(1), 418-429.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment*, *13*(2-3), 108-131.

Spada, N., Shiu, J. L., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning, 65(3),* 723-751.

Stansfield, C. W., Gao, J., & Rivers, W. P. (2010). A concurrent validity study of self-assessments and the federal Interagency Language Roundtable oral proficiency interview. *Russian Language Journal/Русский язык*, *60*, 299-315.

Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal*, *76*(2), 129-141.

Stevenson, D. K. (1981). Beyond faith and face validity: The multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 37-61). TESOL.

Stevenson, D. K. (1985). Authenticity, validity and a tea party. *Language Testing*, *2*(1), 41-47.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from the test scores. *Language Testing, 18*(3), 275-302.

Teng, L. S., Sun, P., & Xu, L. (2018). Conceptualizing writing self-efficacy in English as a foreign language contexts: Scale validation through structural equation modelling. *TESOL Quarterly*, *52*(4), 911-942. Doi: [10.1002/tesq.432](https://doi.org/10.1002/tesq.432)

Teng, L. S., & Zhang, L. J. (2016). A questionnaire-based validation of multidimensional models of self-regulated learning strategies. *Modern Language Journal, 100*(3), 674-701. Doi: 10.1111/modl.12339

Teng, L. S., & Zhang, L. J. (2016). Fostering strategic learning: The development and validation of the writing strategies for motivational regulation questionnaire (WSMRQ). *Asia-Pacific Education Researcher, 25,* 123-134. Doi: [10.1007/s40299-015-0243-4](https://doi.org/10.1007/s40299-015-0243-4)

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, *56*(2), 197-208.

Thorndike, R. L. (Ed.). (1971*). Educational measurement* (2nd Ed.). American Council on Education

Tigchelaar, M., Bowles, R. P., Winke, P., & Gass, S. (2017). Assessing the validity of ACTFL can-do statements for spoken proficiency: A Rasch analysis. *Foreign Language Annals, 50*(3), 584-600.

Ueki, M., & Takeuchi, O. (2012). Validating the L2 motivational self system in a Japanese EFL context: The interplay of L2 motivation, L2 anxiety, self-efficacy, and the perceived amount of information. *Language Education & Technology*, *49*, 1-22.

Underhill, N. (1983). Commonsense in oral testing: Reliability, validity and affective factors. In M. A. Clarke, & J. Handscombe (Eds.), *On TESOL ’82: Pacific perspectives on language learning and teaching* (pp. 125-139). TESOL.

Urmston, A., **Raquel, M., & Tsang, C.** (2103). **Diagnostic testing of Hong Kong tertiary students’ English language proficiency: The development and validation of DELTA.** *Hong Kong Journal of Applied Linguistics 14*(2), 60-82.

Van der Walt, J. L., & Steyn Jr, H. S. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics*, *38*, 191-204.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411-440.

Van Moere, A. (2007). Validity evidence in a group oral test. *Language Testing, 23*, 411-440.

Wainer, H., & Braun, H. I. (2013). *Test validity*. Routledge.

Wang, C., & Bai, B. (2017). Validating the instruments to measure ESL/EFL learners’ self-efficacy beliefs and self-regulated learning strategies. *TESOL Quarterly*, *51*(4), 931-947.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*, 335-353.

Weigle, S.C. (2011).  *Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. TOEFL iBT Research Report TOEFL iBT-15*. Educational Testing Service.

Weinrott, M. R., Jones, R. R., & Bolder, G. R. (1981). Convergent and discriminant validity of five classroom observation systems: A secondary analysis. *Journal of Educational Psychology, 73*(5), 671-680.

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

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

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, *75*(3), 200-208.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1-24.

William, D. (1996). National curriculum assessments and programs of study: Validity and impact. *British Educational Research Journal, 22*(1), 129-141.

Williamson, M. M., & Huot, B. (Eds.), (1993). *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Hampton Press.

Wilson, K. M., & Graves, K. (1999). *Validity of the secondary level English proficiency test at Temple University Japan* (ETS RR-99-11). Educational Testing Service.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161-192.

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers’ perceptions matter. *TESOL Quarterly, 45*(4), 628-660.

Winke, P., Lee, S., Ahn, J. I., Choi, I., & Yoon, H. J. (2018). The cognitive validity of child English language tests: What young language learners and their native-speaking peers can reveal. *TESOL Quarterly, 52*(2), 274-303.

Wintergest, A. C., DeCapua, A., & Itzen, R. C. (2001). The construct validity of one learning styles instrument. *System, 29*(3), 385-404.

Wise, S. L. (1985). The development and validation of a scale measuring attitudes towards statistics. *Educational and Psychological Measurement, 45*(2), 401-405.

Wright, K. L., Hodges, T. S., & McTigue, E. M. (2019). A validation program for the Self-Beliefs, Writing-Beliefs, and Attitude Survey: A measure of adolescents' motivation toward writing. *Assessing Writing, 39*, 64-78. Doi: [10.1016/j.asw.2018.12.004](https://doi.org/10.1016/j.asw.2018.12.004)

Wu, R. Y. F. (2014). *Validating second language reading examinations: Establishing the validity of the GEPT through alignment with the Common European Framework of Reference*. Cambridge University Press.

Xi, X. (2007). Validating TOEFL iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly*, *4*(4), 318–351.

Yan, X., & Staples, S. (2020). Fitting MD analysis in an argument-based validity framework for writing assessment: Explanation and generalization inferences for the ECPE. *Language Testing*, *37*(2), 189–214. <https://doi.org/10.1177/0265532219876226>

Yoo, H., Manna, V. F. (2015). Measuring English language workplace proficiency across subgroups: Using CFA models to validate test score interpretation. *Language Testing, 34*(1), 101–126.

Youn, S. J., & Burch, A. R. (2020). Where conversation analysis meets language assessment: Toward expanding epistemologies and validity evidence. *Papers in Language Testing and Assessment*, *9*(1), iii-xvii.

Young, J. W. (2009). A framework for test validity research on content assessments taken by English language learners. *Educational Assessment*, *14*(3-4), 122-138.

Yun, J. H., Lee, M. B., & Park, Y. H. (2012). A validation study of the language effects of directions, questions, and answer choices on 2014 College Scholastic Ability Test (CSAT). *Korean Journal of Applied Linguistics, 28*(1), 59-86.

Xie, Q. (2011). Is test taker perception of assessment related to construct validity? *International Journal of Testing, 11*(4), 324-348.

Zhao, C. G. (2016). The role of voice in L2 argumentative writing: The development and validation of an analytic rubric. In M. A. Christison & N. Saville (Eds.), *Advancing the field of language assessment: Papers from TIRF doctoral dissertation grantees. Studies in Language Testing 46* (pp. 24-56). Cambridge University Press.

Zhu, Z. (2016). A validation framework for the national English proficiency scale of China. *China Examinations*, *2016*(8), 3-13.