

Seznam literatury

- AERA, APA, & NCME (1999, 2014). *Standards for Educational and Psychological Testing*.
- Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A New Decade of Language Testing: Collaboration and cooperation* (s. 46–57). Ann Arbor, MI: University of Michigan.
- Alderson, Ch. J., Figueras, N., Kuijper, H., Nold, G., Takala, & S. Tardieu, C. (2006). Analysing test of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Assessment Quarterly*, 3(1), 3–30. https://doi.org/10.1207/s15434311laq0301_2
- Anýžová, P. (2013). Ekvivalence položek v mezinárodních datech: základní vymezení a možnosti analýzy. *Data a výzkum – SDA Info* 2013, 7(1), 29–56. <https://doi.org/10.13060/1802-8152.2013.7.1.2>
- Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*. Cambridge: Cambridge University Press.
- Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511667350>
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1
- Bachman, L. F. (1990). *Fundamental Consideration in Language Testing*. Oxford: Oxford University Press.
- Bachman, L. F. (2012). Justifying the use of language assessments: linking interpretations with consequences. Conference paper. Dostupné z: <http://www.sti.chula.ac.th/conference>
- Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.
- Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, 13(2), 125–150. <https://doi.org/10.1177/026553229601300201>

- Bachman, L. F., & Cohen, A. D. (1998). Language testing-SLA interfaces: An update. In L. F. Bachman & A. D. Cohen (Eds.) *Interfaces between SLA and Language Testing Research* (s. 1–31). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139524711.003>
- Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice: Developing Language Tests and Justifying Their Use in the Real World*. Oxford: Oxford University Press.
- Baghaei, P. (2010). Test score equating and fairness in language assessment. *Journal of English Language Studies*, 1(3), 113–128.
- Becker, A. (2016). L2 students' performance on listening comprehension items targeting local and global information. *Journal of English for Academic Purposes*, 24, 1–13. <https://doi.org/10.1016/j.jeap.2016.07.004>
- Bialosiewicz, A., Murphy, K., & Berry, T. (2013). An Introduction to Measurement Invariance Testing: Resource Packet for Participants. CEC. Dostupné z: <http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8>
- Brown, J. D., & Hudson, T. (2002). *Criterion-Referenced Language Testing*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139524803>
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6), 551–558. [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*.
- Council of Europe. (2011). *Manual for Language Test Development and Examining*. Dostupné z: http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf
- Costa-Santos, C., Bernardes, J., Ayres-de-Campos, D., Costa, A., & Costa, C. (2011). The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *Journal of Clinical Epidemiology*, 64, 264–269. <https://doi.org/10.1016/j.jclinepi.2009.11.010>
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306. <https://doi.org/10.1111/j.1745-3984.2000.tb01088.x>
- Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4), 227–246. <https://doi.org/10.1177/0146621604265031>
- Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS RR-10-29*. Dostupné z: <https://www.ets.org/Media/Research/pdf/RR-10-29.pdf>. <https://doi.org/10.1002/j.2333-8504.2010.tb02236.x>

- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington DC: National Academy Press.
- Field, J. (2009). *Listening in the Language Classroom*. Cambridge: Cambridge University. <https://doi.org/10.1017/CBO9780511575945>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Geranpayeh, A. (1994). Are score comparisons across language proficiency test batteries justified? An IELTS-TOEFL comparability study. *Edinburgh Working Paper in Applied Linguistics*, 5, 50–65.
- Geranpayeh, A., & Taylor, L. (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*. Studies in Language Testing 35. Cambridge: UCLES/ CUP
- Goh, Ch. C. M., & Aryadoust, V. (2015). Examining the notion of listening subskill divisibility and its implications for second language listening. *International Journal of Listening*, 29(3), 109–133. <https://doi.org/10.1080/10904018.2014.936119>
- Gwet, K. L. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods For Inter-Rater Reliability Assessment*, 2. Dostupné z: www.agreestat.com
- Gwet, K. L. (2008a). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61, 29–48. <https://doi.org/10.1348/000711006X126600>
- Gwet, K. L. (2008b). Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika*, 73, 407–430. <https://doi.org/10.1007/s11336-007-9054-8>
- Gwet, K. L. (2011). On the Krippendorff's alpha coefficient. Dostupné z: www.agreestat.com
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg: Advanced Analytics.
- Gwet, K. L. (2015). Standard error of Krippendorff's alpha coefficient. Dostupné z: <http://inter-rater-reliability.blogspot.de/2015/08/standard-error-of-krippendorffs-alpha.html>
- Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement*, 76, 609–637. <https://doi.org/10.1177/0013164415596420>
- Harman, H. H. (1976). *Modern Factor Analysis*. (3rd ed.). Chicago: University of Chicago Press.

- Haupt, G., Koch, E. (2012). The argument for evaluating language tests for equivalence across language groups. *Southern African Linguistics and Language Studies*, 30(1), 65–76. <https://doi.org/10.2989/16073614.2012.693715>
- Hendl, J. 2009. *Přehled statistických metod: analýza a metaanalýza dat*. Praha: Portál.
- Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (s. 5–30). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-0-387-49771-6_2
- Holland P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., s. 187–220). Westport, CT: Praeger.
- Chen, F., Huang, X., & MacGregor, D. (2009). Equating or linking: basic concepts and a case study. Presentation originally presented at CAL, Washington. Dostupné z: <https://fliphtml5.com/xrgx/bfuj/basic>
- Choi, I., Sung, K., & Boo, J. (2003). Comparability of a paper-based language tests and a computer-based language test. *Language Testing*, 20(3), 295–320. <https://doi.org/10.1191/0265532203lt258oa>
- Chráska, M. (2007). *Metody pedagogického výzkumu. Základy kvantitativního výzkumu*. Praha: Grada
- Chvál, M., Straková, J., & Procházková, I. (2015). *Hodnocení výsledků vzdělávání didaktickými testy*. Praha: Česká školní inspekce.
- Jelínek, M., Květoň, P., & Vobořil, D. (2011). *Teorie odpovědi na položku a počítačové adaptivní testování*. Praha: Grada.
- Kolen, M. J., & Brennan, R. L. (2004, 2014). *Test equating, scaling, and linking: Methods and practice*. New York: Springer. <https://doi.org/10.1007/978-1-4757-4310-4>
- Kirkeboen, G. (2009). Decision behaviour – Improving expert judgement. Dostupné : http://www.concept.ntnu.no/attachments/058_Kirkebooen%20%20-%20Expert%20judgement.pdf
- Khalifa, H., Weir, C. (2009). *Examining reading*. Cambridge: Cambridge University Press.
- Klein, D. (2018). Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal*, 18(4), 871–901. <https://doi.org/10.1177/1536867X1801800408>
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag. <https://doi.org/10.1007/978-1-4757-4310-4>
- Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology*, 64(6), 701–702. <https://doi.org/10.1016/j.jclinepi.2010.12.001>

- Krippendorff, K. (2004). *Content analysis. An introduction to its methodology*. Thousand Oaks, CA: Sage Publications, Inc.
- Kunnan, A. J., & Carr, N. (2017). A comparability study between the general English proficiency test-advanced and the internet-based test of English as a foreign language. *Language Testing in Asia*, 7(17). <https://doi.org/10.1186/s40468-017-0048-x>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. <https://doi.org/10.2307/2529310>
- Lavrakas, P. J. (Ed.). (2008). *Encyclopedia of survey research methods*, Thousand Oaks, CA: Sage Publications, Inc. <https://doi.org/10.4135/9781412963947>
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83–102. https://doi.org/10.1207/s15324818ame0601_5
- Livingston, S. A. (2004). *Test score equating (without IRT)*. Educational Testing Service. Dostupné z: www.ets.org
- Lumley, T. (1993). Reading comprehension sub-skills: teachers' perceptions of content in an EAP test. *Melbourne Papers in Language Testing*, 2(1), 24–57.
- McCray, G. (2013). Assessing inter-rater agreement for nominal judgement variables. Paper presented at the *Language Testing Forum*. Nottingham, November 15-17, 2013. Dostupné z: www.agreestat.com
- Messick, S. (1987). Validity. *ETS Research Report Series*. doi:10.1002/j.2330-8516.1987.tb00244.x
- Messick, S. (1993). Foundations of validity: meaning and consequences in psychological assessment. *ETS Research Report Series*. doi:10.1002/j.2333-8504.1993.tb01562.x
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Michaelides, M. P., & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: a bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, 27(1), 46–57. <https://doi.org/10.1080/08957347.2013.853069>
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects (Policy Information Rep.)*. Princeton, NJ: ETS.
- O'Loughlin, K. (1997). *The Comparability of Direct and Semi-direct Speaking Tests: A case study*. Unpublished Ph.D. thesis. Melbourne: University of Melbourne. Dostupné z: <https://minerva-access.unimelb.edu.au/handle/11343/38817>

- Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement*, 28(4), 247–273. <https://doi.org/10.1177/0146621604265033>
- Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511733086>
- Purpura, J. E. (2014a). Assessing grammar. In A. J. Kunnan (Ed.). *Companion to language assessment* (s. 100–124). Oxford: Wiley. <https://doi.org/10.1002/9781118411360.wbcla147>
- Purpura, J. E. (2014b). Cognition in language assessment. In A. J. Kunnan (Ed.), *Companion to language assessment* (s. 1452–1476). Oxford: Wiley. <https://doi.org/10.1002/9781118411360.wbcla150>
- Purpura J. E. (2017) Assessing meaning. In E. Shohamy, I. Or., & S. May (Eds.), *Language Testing and Assessment. Encyclopedia of Language and Education* (s. 33–61). (3rd ed.). Cham: Springer. https://doi.org/10.1007/978-3-319-02261-1_1
- Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning Technology* 5, 38–59.
- Sireci, S., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing* 20(2), 148–166. <https://doi.org/10.1191/0265532203lt249oa>
- Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford: Oxford University Press.
- Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, 41(10), 949–958. [https://doi.org/10.1016/0895-4356\(88\)90031-5](https://doi.org/10.1016/0895-4356(88)90031-5)
- Urbina. S. (2004). *Essentials of Psychological Testing*. New Jersey: John Wiley & Sons, Inc.
- van den Heuvel-Hanhuizen, M., Robitzsch, A., Treffers, A., & Köller, O. (2009). Large-scale assessment of change in student achievement: Dutch primary school students' results on written division in 1997 and 2004 as an example. *Psychometrika*, 74(2), 351–365. <https://doi.org/10.1007/s11336-009-9110-7>
- van de Vijver, F. & Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (s. 39–63). Mahwah, NJ: IEA Lawrence Erlbaum Associates, Publishers.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70. <https://doi.org/10.1177/109442810031002>

- Verhelst N. D., & Glas C. A. W. (1995). The one parameter logistic model. In *Fischer G. H., Molenaar I. W. (Eds.), Rasch Models*. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4230-7_12
- von Davier. A. A. (2011) A statistical perspective on equating test scores. In A. A. von Davier (Ed.). *Statistical models for test equating, scaling, and linking* (s. 1–17). New York, NY: Springer. https://doi.org/10.1007/978-0-387-98138-3_1
- Watson, J. C. (2017) Establishing evidence for internal structure using exploratory factor analysis. *Measurement and Evaluation in Counseling and Development*, 50(4), 232–238. <https://doi.org/10.1080/07481756.2017.1336931>
- Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22, 281–300. <https://doi.org/10.1191/0265532205lt309oa>
- Weir, C. (2005). *Language Testing and Validation. An Evidence-based Approach*. Basingstoke: Palgrave, MacMillan. <https://doi.org/10.1057/9780230514577>
- Weir, C, Wu, R. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing*, 23, 167–197. <https://doi.org/10.1191/0265532206lt326oa>
- Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child development perspectives*, 4(1), 10–18. <https://doi.org/10.1111/j.1750-8606.2009.00110.x>
- Wu, R. Y. (2014). *Validating Second Language Reading Examinations Establishing the validity of the GEPT through alignment with the Common European Framework of Reference*. Studies in Language Testing. Cambridge: Cambridge University Press.