UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR PUSTAKA

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2-3), 162-172. https://doi.org/10.1016/j.stueduc.2005.05.008

Adams, R., & Wu, M. (Ed.). (2002). *PISA 2000 technical report.* OECD. https://www.oecd.org/pisa/data/33688233.pdf

Adams, R., Cloney, D., Wu, M., Osses, A., & Vista, A. (2020). *ConQuest manual.* ACER Press.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23. https://doi.org/10.1177/0146621697211001

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). Macmillan.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report.* National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main1998/2001509.pdf

American Psychological Association. (2023). *APA guidelines for the undergraduate psychology major* (Version 3.0). American Psychological Association. https://www.apa.org/about/policy/undergraduate-psychology-major.pdf

American Psychological Association. (2015). *APA dictionary of psychology* (2nd ed.). American Psychological Association.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education.

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

(2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*(1), 69–81. https://doi.org/10.1007/BF02293746

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Andrich, D. (1988). *Rasch models for measurement.* Sage Publications, Inc.

Andrich, D., & Marais, I. (2019). *A course in rasch measurement theory: measuring in the educational, social and health sciences*. Springer Nature Singapore Pte Ltd.

Andrich, D. A. (2011). Rating scales and Rasch measurement. *Expert Review of. Pharmacoeconomics & Outcomes Research,* 11(5), 571–585. https://doi.org/10.1586/erp.11.59

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-597). American Council on Education.

AP2TPI. (2018). *Perubahan atas surat keputusan AP2TPI No. 01/kep/AP2TPI/2015 tentang kurikulum inti program studi psikologi jenjang sarjana.* http://ap2tpi.or.id/wp-content/uploads/2019/05/SK-AP2TPI-Perubahan-Kurikulum-Inti-Program-Studi-Sarjana-Final-sdh-ttd-22-november-2018.pdf

Asparouhov, T., & Muthen, B. (2010). *Plausible values for latent variables using Mplus.* Muthen & Muthen.

Azwar, S. (2021). *Penyusunan skala psikologi* (3rd ed.). Pustaka Pelajar.

Azwar, S. (2024). *Validitas prediktif diferensial.* Pustaka Pelajar.

Azwar, S. (2025). *Dasar-dasar psikometrika* (3rd ed.). Pustaka Pelajar.

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests.* Oxford University Press.

Bai, Y., Liang, H., Qi, C., & Zuo, S. (2023). An assessment of eighth graders' mathematics higher order thinking skills in the Chinese context. *Canadian Journal of Science, Mathematics and Technology Education, 23*(2), 365-382. https://doi.org/10.1007/s42330-023-00279-w

Ben-Zvi, D., & Garfield, J. (2004). *The challenge of developing statistical literacy, reasoning and thinking.* Springer.

Berk, R. A. (1976). Determination of optional cutting scores in criterion-referenced measurement. *The Journal of Experimental Education, 45*(2), 4-9. https://doi.org/10.1080/00220973.1976.11011567

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137–172. https://doi.org/10.2307/1170289

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association, 39*(227), 357-365. https://doi.org/10.1080/01621459.1944.10500699

Berndt, M., Schmidt, F. M., Sailer, M., Fischer, F., Fischer, M. R., & Zottmann, J. M. (2021). Investigating statistical literacy and scientific reasoning & argumentation in medical-, social sciences-, and economics students. *Learning and Individual Differences, 86*, 101963. https://doi.org/10.1016/j.lindif.2020.101963

Bhatti, H. A., Mehta, S., McNeil, R., Yao, S.-Y., & Wilson, M. (2023). A scientific approach to assessment: Rasch measurement and the four building blocks. In X. Liu & W. J. Boone (Eds.), *Advances in*

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

applications of Rasch measurement in science education (pp. 157-187). Springer. https://doi.org/10.1007/978-3-031-28776-3_7

Blomeke, S., & Gustafsson, J.-E. (2017). *Standard setting in education: The Nordic countries in an international perspective.* Springer.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals* (1st ed.). Longman Group.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*(4), 443-459. https://doi.org/10.1007/BF02293801

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197-211. https://doi.org/10.1111/j.1745-3984.1997.tb00515.x

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: fundamental measurement in the human sciences* (4th ed.). Routledge.

Boone, W. J. (2020). Rasch basics for the novice. In M. S. Khine (Ed.), *Rasch measurement: Applications in quantitative educational research* (pp. 9-30). Springer.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, *4*(2), 219–240. https://doi.org/10.1177/014662168000400209

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., & Harris, E. L. (1995). Describing performance standards: validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3*(1), 9-51. https://doi.org/10.1207/s15326977ea0301_2

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Cazden, C. B. (1982). Contexts for literacy: In the mind and in the classroom. *Journal of Literacy Research, 14*(4), 413-427. https://doi.org/10.1080/10862968209547467

Cheng, J., Zheng, X., & Zhu, Y. (2021). Chinese eighth graders' competencies in mathematical reasoning. In B. Xu, Y. Zhu, & X. Lu (Eds.), *Beyond shanghai and PISA: Cognitive and non-cognitive competencies of Chinese students in mathematics* (pp. 187-207). Springer. https://doi.org/10.1007/978-3-030-68157-9_11

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31*(6), 467–482. https://doi.org/10.1177/0146621606292933

Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2009). Multidimensional rasch analysis of a psychological test with multiple subtest: a statistical solution for the bandwidth-fidelity dilemma. *Educational and Psychological Measurement, 69*(3), 369-388. https://doi.org/10.1177/0013164408323241

Cho, E. (2023). Interchangeability between factor analysis, logistic IRT, and normal ogive IRT. *Frontiers in Psychology, 14*, 1267219. https://doi.org/10.3389/fpsyg.2023.1267219

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178-194. https://doi.org/10.1177/0146621616677520

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30(2), 93–106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2012). An introduction to contemporary standard setting: concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: foundations, methods, and innovations* (2nd ed., pp. 3-14). Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications Ltd.

Cohen, B. H. (2013). *Explaining psychological statistics* (4th ed.). John Wiley & Sons, Inc.

Cook, R. M., & Wind, S. A. (2024). Item response theory: A modern measurement approach to reliability and precision for counseling researchers. *Measurement and Evaluation in Counseling and Development, 57*(2), 116–135. https://doi.org/10.1080/07481756.2023.2301284

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper.

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd ed.). Guilford Press.

Draney, K., & Kennedy, C. (2010). The standard setting criterion-mapping method. In C. Harsch, H. A. Pant, & O. Koller (Eds.). *Calibrating standards-based assessment tasks for English as a first foreign language: Vol. 2. Standard-setting procedures in Germany* (pp. 75–88). Waxmann Verlag GmbH.

Draney, K., & Wilson, M. (2009). Selecting cut scores with a composite of item types: The Construct Mapping procedure. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to*

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

score reporting using Rasch measurement models (pp. 276–293). JAM Press.

Dunham, B., Yapa, G., & Yu, E. (2015). Calibrating the difficulty of an assessment tool: The Blooming of a statistics examination. *Journal of Statistics Education, 23*(3). https://doi.org/10.1080/10691898.2015.11889745

Educational Testing Service. (2014). *TOEFL ITP test score descriptors.* Educational Testing Service. https://www.ets.org/s/toefl_itp/pdf/test_score_descriptors.pdf

Educational Testing Service. (2021). *TOEFL iBT performance descriptors.* Educational Testing Service. https://www.ets.org/s/toefl/pdf/pd-toefl-ibt.pdf

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika, 75*(3), 474–497. https://doi.org/10.1007/s11336-010-9161-9

Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A., & Hagell, P. (2022). Transformation of Rasch model logits for enhanced interpretability. *BMC Medical Research Methodology, 22*(1), 332. https://doi.org/10.1186/s12874-022-01816-1

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50*(1), 195-212. https://doi.org/10.3758/s13428-017-0862-1

Feuerstahler, L., & Wilson, M. (2019). Scale alignment in between-item multidimensional rasch models. *Journal of Educational Measurement, 56*(2), 280-301. https://doi.org/10.1111/jedm.12209

Feuerstahler, L., & Wilson, M. (2021). Scale alignment in the between-item multidimensional partial credit model. *Applied Psychological*

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Measurement, 45(4), 268–282. https://doi.org/10.1177/01466216211013103

Freund, R. L. (2019). *Rasch and rationality: scale typologies as applied to item response theory.* Doctoral dissertation, University of California, Berkeley.

Gal, I. (2002). Adults' statistical literacy: meanings, components, responsibilites. *International Statistical Review, 70*(1), 1-25. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Gal, I. (2004). Statistical literacy. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 47-78). Springer.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2(*1), 22-38. https://doi.org/10.52041/serj.v2i1.557

Garfield, J. (2011). Statistical literacy, reasoning, and thinking. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science.* Springer.

Garfield, J., delMas, R., & Zieffler, A. (2010). Assessing statistical thinking. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 175–186). John Wiley & Sons.

Gebhardt, E. (2016). *Latent path models within an IRT framework.* Doctoral dissertation, University of Melbourne

Geisinger, K. F. (2021). The history of norm- and criterion-referenced testing. In B. E. Clauser, & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 42-64). Routledge.

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. *American Psychologist, 18*(8), 519-521. https://doi.org/10.1037/h0049294

Gonulal, T. (2016). *Statistical literacy among second language acquisition graduate students.* Doctoral dissertation, Michigan State University

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22-25. https://doi.org/10.52041/serj.v16i1.209

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450. https://doi.org/10.1037/1082-989X.6.4.430

Grosse, M. E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation and the Health Professions*, *9*(3), 267-285. https://doi.org/10.1177/016327878600900301

Hambleton, R. K. (1978). On the use of cut-off scores with criterion referenced tests in instructional settings. *Journal of Educational Measurement, 15*(4), 277-290. https://doi.org/10.1111/j.1745-3984.1978.tb00075.x

Hambleton, R. K. (1994). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice*, 13(4), 21–26. https://doi.org/10.1111/j.1745-3992.1994.tb00567.x

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 479–

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

494). American Psychological Association. https://doi.org/10.1037/14049-023

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessment. *Applied Psychological Measurement, 24*(4), 355-366. https://doi.org/10.1177/01466210022031804

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: a review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47. https://doi.org/10.3102/00346543048001001

Harrell-Williams, L. M., & Wolfe, E. W. (2013). The influence of between-dimension correlation, misfit, and test length on multidimensional rasch model information-based fit index accuracy. *Educational and Psychological Measurement, 73*(4), 672-689. https://doi.org/10.1177/0013164413477654

Harris, K. M., Wang, L., Mu, G. M., Lu, Y., So, C., Zhang, W., ... & Ho, R. C. (2023). Measuring the suicidal mind: The 'open source'Suicidality Scale, for adolescents and adults. *PloS one, 18*(2), e0282009. https://doi.org/10.1371/journal.pone.0282009

Hayat, B., & Putra, M. D. K. (2025). *Aplikasi Rasch model dengan program RStudio*. Inparametric Press.

Hayat, B., Putra, M. D. K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions. *Jurnal Penelitian dan Evaluasi Pendidikan*, *24*(1), 39-50. https://doi.org/10.21831/pep.v24i1.29871

Hayat, B., Putra, M. D. K., Rahmawati., & Isa, K. (2023). Rasch calibration and Differential Item Functioning (DIF) analysis of the Indonesian National Assessment Program-Language (INAP-L). *World Journal*

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

of English Language, 13(6), 402-413. https://doi.org/10.5430/wjel.v13n6p402

Heidary, S., Abdollahi, A., Jarrahi, E., & Ahmed, A. A. A. (2022). Psychometric Assessment of the Persian Version of Compassion Scale for Adolescents. *Evaluation & the Health Professions, 45*(3), 297-302. https://doi.org/10.1177/01632787211053855

Hoijtink, H., Rooks, G., & Wilmink, F. W. (1999). Confirmatory factor analysis of items with a dichotomous response format using the multidimensional Rasch model. *Psychological Methods, 4*(3), 300–314. https://doi.org/10.1037/1082-989X.4.3.300

Holman, R., Lindeboom, R., Glas, C. A. W., Vermeulen, M., & de Haan, R. J. (2003). Constructing an item bank using item response theory: The AMC linear disability score project. *Health Services & Outcomes Research Methodology, 4*(1), 19-33. https://doi.org/10.1023/A:1025824810390

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubley, A. M. (2021). Response processes validity evidence: Understanding the meaning of scores from psychological measures. In P. Graf & D. J. A. Dozois (Eds.), *Handbook on the state of the art in applied psychology* (pp. 413–434). Wiley Blackwell.

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in*

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

industrial and organizational psychology (pp. 3–19). American Psychological Association. https://doi.org/10.1037/14047-001

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19–20. https://doi.org/10.1111/j.1745-3992.2006.00053.x

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). American Council on Education.

Johannssen, A., Chukhrova, N., Schmal, F., & Stabenow, K. (2021). Statistical literacy—misuse of statistics and its consequences. *Journal of Statistics and Data Science Education, 29*(1), 54-62. https://doi.org/10.1080/10691898.2020.1860727

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1–27). Elsevier.

Kane, M. T. (1987). On the use of IRT models with judgmental standard setting procedures. *Journal of Educational Measurement*, *24*(4), 333- 345. https://doi.org/10.1111/j.1745-3984.1987.tb00284.x

Kane, M. (1994). Validating the performance standards associates with passing scores. *Review of Educational Research, 64*(3), 425-461. https://doi.org/10.3102/00346543064003425

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). SAGE Publications, Inc. https://doi.org/10.4135/9781452226576

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*(2), 152–176.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting: a literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12. https://doi.org/10.1111/j.1745-3992.2006.00047.x

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement*, *20*(2), 155–168. https://doi.org/10.1177/014662169602000205

Khorramdel, L., von Davier, M., Gonzalez, E., & Yamamoto, K. (2020). Plausible values: principles of item response theory and multiple imputations. In D. B. Maehler, & B. Rammstedt. (Eds.), *Large-scale cognitive assessment: analyzing PIAAC data* (pp. 27-48). Springer.

Kilani, H., Markov, I. V., Francis, D., & Grigorenko, E. L. (2024). Screens and preschools: The Bilingual English Language Learner Assessment as a curriculum-compliant digital application. *Children, 11*(8), 914. https://doi.org/10.3390/children11080914

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*(4), 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science + Business Media. https://doi.org/10.1007/978-1-4939-0317-7

Koller, I., Levenson, M. R., & Glück, J. (2017). What do you think you are measuring? A mixed-methods procedure for assessing the content validity of test items and theory-based scaling. *Frontiers in Psychology, 8*: 126. https://doi.org/10.3389/fpsyg.2017.00126

Kreiner, S. (2013). The Rasch model for dichotomous items. In K. B. Christensen, S. Kreiner, & M. Mesbah. (Eds.), *Rasch models in health*. John Wiley & Sons, Inc.

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Kurnia, A. B., Lowrie, T., & Patahuddin, S. M. (2024). The development of high school students' statistical literacy across grade level. *Mathematics Education Research Journal, 36*(Suppl 1), 7-35. https://doi.org/10.1007/s13394-023-00449-x

L'Boy, D., & Khan, R. N. (2023). A Rasch-model-based hierarchical framework for statistical literacy and learning. *International Journal of Mathematical Education in Science and Technology, 54*(9), 1874-1887. https://doi.org/10.1080/0020739X.2023.2261453

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51*(2-3), 220–239. https://doi.org/10.1080/00273171.2015.1134306

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lewis, D., & Lord-Bessen, J. (2018). Standard setting. In W. J. van der Linden (Ed.), *Handbook of item response theory, Vol. 3 Application* (pp. 229-248). CRC Press.

Linacre, J. M. (1994). *Many-facet rasch measurement* (2nd ed.). MESA Press.

Linacre, J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement, 7*(1), 1-11.

Linacre, J. M. (2010). Two perspectives on the application of Rasch models. *European Journal of Physical and Rehabilitation Medicine*, *46*, 301-302.

Linacre, J. M. (2013). Reliability, separation and strata: Percentage of sample in each level. *Rasch Measurement Transactions, 26*(4), 1399. https://www.rasch.org/rmt/rmt264g.htm

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Luecht, R. M. (2020). Generating performance-level descriptors under a principled assessment design paradigm: An example for assessments under the Next-Generation Science Standards. *Educational Measurement: Issues and Practice, 39*(4), 105-115. https://doi.org/10.1111/emip.12356

Luecht, R. M., de Champlain, A., & Nungester, R. J. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education, 3*(1), 29-41. https://doi.org/10.1023/A:1009789314011

Luo, Y., & Dimitrov, D. M. (2019). A short note on obtaining point estimates of the IRT ability parameter with MCMC estimation in Mplus: how many plausible values are needed?. *Educational and Psychological Measurement, 79*(2), 272-287. https://doi.org/10.1177/0013164418777569

MacCann, R. G. (2009). Standard setting with dichotomous and constructed response items: some Rasch model approaches. *Journal of Applied Measurement, 10*(4), 438-454.

Mair, P. (2018). *Modern psychometrics with R*. Springer International Publishing AG.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/BF02296272

Maydeu-Olivares, A., & Montaño, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, *78*(1), 116–133. https://doi.org/10.1007/s11336-012-9293-1

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287-2305. https://doi.org/10.3758/s13428-020-01398-0

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

McVee, M. B., Schucker, K. A., Silvestri, K. N., & Tripp, J. N. (2023). Using multiliteracies and multimodality to explore engineering literacies and problem-scoping for multilingual learners: A perspective on multiliteracies for STEM learning. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education* (4th ed, pp. 50-63). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.13039-8

Miller, M., & Veatch, N. (2010). Teaching literacy in context: Choosing and using instructional strategies. *The Reading Teacher, 64*(3), 154-165. https://doi.org/10.1598/RT.64.3.1

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, *43*(2), 205–216. https://doi.org/10.2307/1169937

Mislevy R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Lawrence Erlbaum Associates Publishers.

Mokkink, L., Herbelet, S., Tuinman, P., & Terwee, C. (2025). Content validity: judging the relevance, comprehensiveness and comprehensibility of an outcome measurement instrument–a COSMIN perspective. *Journal of Clinical Epidemiology, 185*: 111879. https://doi.org/10.1016/j.jclinepi.2025.111879

Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

understand the structure of matter. *Journal of Research in Science Teaching, 54*(8), 1024-1048. https://doi.org/10.1002/tea.21397

Mullis, I. V. S., & Martin, M. O. (Ed.). (2017). *TIMSS 2019 assessment frameworks*. International Association for the Evaluation of Educational Achievement (IEA).

Muthén, L. K., & Muthén, B. O. (1998–2020). *Mplus user's guide: Statistical analysis with latent variables* (8th ed.). Muthén & Muthén.

Nason, K., & DeMars, C. (2025). Another Look at Yen's Q3: Is. 2 an Appropriate Cut-Off? *Journal of Educational Measurement, 62*(2), 345-359. https://doi.org/10.1111/jedm.12432

Nitko, A. J. (1980). Distinguishing the many varieties of criterion-referenced tests. *Review of Educational Research*, 50(3), 461–485. https://doi.org/10.2307/1170441

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445-465. https://doi.org/10.1016/0346-251X(95)00032-F

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing,* 15(2), 217-262. https://doi.org/10.1177/026553229801500204

Nunnally, J. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, 20(4), 641–650. https://doi.org/10.1177/001316446002000401

Ockwell, L., Daraganov, A., & Schulz, W. (2020). Scaling procedures for ICILS 2018 test items. In J. Fraillon, W. Schulz, D. Duckworth, J. Ainley, & T. Friedman (Eds.), *IEA international computer an information literacy study 2018 technical report* (pp. 133-158). International Association for the Evaluation of Educational Achievement (IEA).

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

OECD. (2014). *PISA 2012 technical report.* OECD Publishing.

OECD. (2018). *PISA 2022 mathematics framework (draft).* OECD Publishing. https://pisa2022-maths.oecd.org/files/PISA%202022%20Mathematics%20Framework%20Draft.pdf

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://www.oecd-ilibrary.org/docserver/b25efab8-en.pdf?expires=1638885172&id=id&accname=guest&checksum=1E56F59F704092E641336229BFA61B83

Olsen, R. V., & Nilsen, T. (2017). Standard setting in PISA and TIMSS and how these procedures can be used nationally. In S. Blomeke, & J-E. Gustafsson (Eds.), *Standard setting in education: the nordic countries in an international perspective* (pp. 69-84). Springer International Publishing.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptor for reporting overall examinee performance. *Language Assessment Quarterly, 12*(2), 153-177. https://doi.org/10.1080/15434303.2015.1008480

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178. https://doi.org/10.2307/1165199

Peeters, M. J., & Harpe, S. E. (2020). Updating conceptions of validity and reliability. *Research in Social & Administrative Pharmacy, 16*(8), 1127–1130. https://doi.org/10.1016/j.sapharm.2019.11.017

Phadke, S. (2022). *Measurement, assessment, and improvement of statistical literacy in relevant contexts*. Doctoral dissertation, Pennsylvania State University

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Phadke, S., Beckman, M., & Morgan, K. L. (2024). Examining the role of context in statistical literacy assessment. *Statistics Education Research Journal, 23*(1), 1–17. https://doi.org/10.52041/serj.v23i1.529

Perlman, B., & McCann, L. I. (1999a). The most frequently listed courses in the undergraduate psychology curriculum. *Teaching of Psychology*, *26*(3), 177–182. https://doi.org/10.1207/S15328023TOP260303

Perlman, B., & McCann, L. I. (1999b). The structure of the psychology undergraduate curriculum. *Teaching of Psychology*, *26*(3), 171–176. https://doi.org/10.1207/S15328023TOP260302

Philpot, R., Lindquist, M., Mullis, I. V. S., & Aldrich, C. E. A. (2021). TIMSS 2023 mathematics framework. In I. V. S. Mullis, M. O. Martin, & M. von Davier (Eds.), *TIMSS 2023 assessment frameworks* (pp. 5-18). International Association for the Evaluation of Educational Achievement (IEA).

Piatti, A., Adorni, G., El-Hamamsy, L., Negrini, L., Assaf, D., Gambardella, L., & Mondada, F. (2022). The CT-cube: A framework for the design and the assessment of computational thinking activities. *Computers in Human Behavior Reports, 5*, 100166. https://doi.org/10.1016/j.chbr.2021.100166

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467. https://doi.org/10.1002/nur.20199

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 74–91). Cambridge University Press.

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1-9. https://doi.org/10.1111/j.1745-3984.1969.tb00654.x

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 49-57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* Taylor & Francis.

Raykov, T., & Marcoulides, G. A. (2013). *Basic statistics: An introduction with R*. Rowman & Littlefield Publishers, Inc.

Raykov, T., Marcoulides, G. A., & Patelis, T. (2013). Saturated versus just identified models: A note on their distinction. *Educational and Psychological Measurement, 73*(1), 162-168. https://doi.org/10.1177/0013164412457367

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4-18. https://doi.org/10.1111/j.1745-3992.2006.00052

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Reckase, M. D., & Xu, J. R. (2015). The evidence for a subscore structure in a test of english language competency for english language learners. *Educational and Psychological Measurement, 75*(5), 805-825. https://doi.org/10.1177/0013164414554416

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

construct-modeling approach. *Journal of Educational Psychology, 103*(1), 85–104. https://doi.org/10.1037/a0021334

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3), 1-12. https://doi.org/10.1080/10691898.2002.11910678

Sabbag, A. G. (2016). *Examining the relationship between statistical literacy and statistical reasoning.* Doctoral dissertation, University of Minnesota.

Sabbag, A., Garfield, J., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning: The REALI instrument. *Statistics Education Research Journal, 17*(2), 141–160. https://doi.org/10.52041/serj.v17i2.163

Sabbag, A., Zieffler, A., & Ng, C. (2025). Can we distinguish statistical literacy and statistical reasoning? *Statistics Education Research Journal, 24*(1), 1–19. https://doi.org/10.52041/serj.v24i1.587

Salzberger, T. (2010). Does the Rasch model convert an ordinal scale into an interval scale? *Rasch Measurement Transactions, 24*(2), 1273-1275.

Schneider, M. C., Agrimson, J., & Veazey, M. (2022). The relationship between item developer alignment of items to range achievement-level descriptors and item difficulty: Implications for validating intended score interpretations. *Educational Measurement: Issues and Practice, 41*(2), 12–24. https://doi.org/10.1111/emip.12480

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*(4), 347-362. https://doi.org/10.1177/01466219922031464

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Schwartz, R. A. (2012). *The development and psychometric modeling of an embedded assessment for a data modeling and statistical reasoning learning progression.* Unpublished doctoral dissertation, University of California, Berkeley.

Sharma, S. (2017). Definition and models of statistical literacy: a literature review. *Open Review of Educational Research, 4*(1), 118-133. https://doi.org/10.1080/23265507.2017.1354313

She, H. C., Stacey, K., & Schmidt, W. H. (2018). Science and mathematics literacy: PISA for better school education. *International Journal of Science and Mathematics Education, 16*(Suppl 1), 1-5. https://doi.org/10.1007/s10763-018-9911-1

Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*(4), 447-467. https://doi.org/10.1177/014662168000400403

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*(4), 563-576. https://doi.org/10.1007/BF02296196

Slinde, J. A., & Linn, R. L. (1979). The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement, 3*(4), 437-452. https://doi.org/10.1177/014662167900300402

Stowell, J. R., & Addison, W. E. (Eds.). (2017). *Activities for teaching statistics and research methods: A guide for psychology instructors*. American Psychological Association. https://doi.org/10.1037/0000024-000

Sweet, T., & Wang, S. (2025). Network science in psychology. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000745

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 3. Testing and assessment in school psychology and education* (pp. 455–477). American Psychological Association. https://doi.org/10.1037/14049-022

Tennant, A., & Conaghan, P. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care and Research, 5*(8), 1358–1362. https://doi.org/10.1002/art.23108

Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024a). Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model. *Disability and Rehabilitation: An International, Multidisciplinary Journal, 46*(3), 591–603. https://doi.org/10.1080/09638288.2023.2169771

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024b). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation: An International, Multidisciplinary Journal, 46*(3), 604–617. https://doi.org/10.1080/09638288.2023.2169772

Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a Statistic Classroom Exam: Assessing its relation with item cognitive level classified according to Bloom’s taxonomy. *Frontiers in Psychology*, *9*: 1585. https://doi.org/10.3389/fpsyg.2018.01585.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*(2), 175–186. https://doi.org/10.1007/BF02296273

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on test including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39-49. https://doi.org/10.1177/014662169501900105

Tiro, M. A. (2018). *Strategi aksi gerakan nasional literasi statistika di Indonesia*. Artikel dipresentasikan dalam Seminar Nasional VARIANSI ke-2, Makassar, Indonesia, 9-10 Oktober. https://ojs.unm.ac.id/variansistatistika/article/view/7193

Tomcho, T. J., Rice, D., Foels, R., Folmsbee, L., Vladescu, J., Lissman, R., Matulewicz, R., & Bopp, K. (2009). APA's learning objectives for research methods and statistics in practice: a multimethod analysis. *Teaching of Psychology, 36*(2), 84-89. https://doi.org/10.1080/00986280902739693

UIN Syarif Hidayatullah Jakarta. (2019). *Pedoman akademik program strata 1 (S1) 2019-2020.* UIN Syarif Hidayatullah Jakarta.

Umar, J. (1987). *Robustness of the simple linking procedure in item banking using the Rasch model.* Doctoral dissertation, University of California, Los Angeles.

Universitas Brawijaya. (2019). *Konversi mata kuliah program studi strata-1 psikologi*. https://psikologi.ub.ac.id/wp-content/uploads/2019/07/Konversi-Mata-Kuliah-Program-Studi-S1-Psikologi.pdf

van der Linden, W. J. (1982). Criterion-referenced measurement: its main applications, problems and findings. *Evaluation in Education, 5*(2), 97-118. https://doi.org/10.1016/0191-765X(82)90012-X

van Laar, S., & Braeken, J. (2022). Caught off base: A note on the interpretation of incremental fit indices. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(6), 935-943. https://doi.org/10.1080/10705511.2022.2050730

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). *What are plausible values and why are they useful*? IEA-ETS Research Institute.

Walker, H. M. (1951). Statistical literacy in the social sciences. *The American Statistician, 5*(1), 6-12. https://doi.org/10.2307/2685917

Wallman, K. K. (1993). Enchanching statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*(421), 1-8. https://doi.org/10.1080/01621459.1993.10594283

Wang, N. (2003). Use of the Rasch IRT model in standard setting: an item-mapping method. *Journal of Educational Measurement*, *40*(3), 231-253. https://doi.org/10.1111/j.1745-3984.2003.tb01106.x

Wang, W.-C., Yao, G., Tsai, Y.-J., Wang, J.-D., & Hsieh, C.-L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research*, *15*(4), 607–620. https://doi.org/10.1007/s11136-005-4365-7

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Watson, J. M. (1997). Assessing statistical literacy using the media. In I. Gal and J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107-121). IOS Press and The International Statistical Institute.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Lawrence Erlbaum Associates Publishers.

Watson, J., & Callingham, R. (2003). Statistical literacy: a complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46. https://doi.org/10.52041/serj.v2i2.553

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Whitely, S. E., & Dawis, R. V. (1974). The nature objectivity with the Rasch model. *Journal of Educational Measurement, 11*(3), 163-178. https://doi.org/10.1111/j.1745-3984.1974.tb00988.x

Widhiarso, W. (2019). Pembuktian validitas terkait struktur tes potensi akademik pascasarjana (PAPS) Universitas Gadjah Mada. *Jurnal Psikologi, 46*(2), 145-162. https://doi.org/10.22146/jpsi.38223

Widhiarso, W., & Ridho, A. (2022). *Validation of setting and design of multi-stage testing (MST) to portray students' achievement on reading literacy in AKMI 2021*. In Proceedings of the International Conference on Madrasah Reform 2021 (ICMR 2021). https://doi.org/10.2991/assehr.k.220104.002

Wilson, M. (2013). Seeking a balance between the statistical and scientific elements in psychometrics. *Psychometrika*, *78*(2), 211–236. https://doi.org/10.1007/s11336-013-9327-3

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). Routledge. https://doi.org/10.4324/9781003286929

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (pp. 325-332). Springer-Verlag.

Wilson, M., & Gochyyev, P. (2020). Having your cake and eating it too: Multiple dimensions and a composite. *Measurement, 151*, 107247. https://doi.org/10.1016/j.measurement.2019.107247

Wilson, M., & Santelices, M. V. (2017). Weaknesses of the traditional view of standard setting and suggested alternative. In S. Blomeke, & J-E. Gustafsson (Eds.), *Standard setting in education: the nordic countries in an international perspective* (pp. 31-48). Springer International Publishing.

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208. https://doi.org/10.1207/S15324818AME1302_4

Wongvorachan, T., Srisuttiyakorn, S., & Sriklaub, K. (2024). Optimizing learning: Predicting research competency via statistical proficiency. *Trends in Higher Education, 3*(3), 540-559. https://doi.org/10.3390/higheredu3030032

Wright, B. D. (1968). *Sample-free test calibration and person measurement.* In Proceedings of the 1967 Invitational Conference of Testing Problems. Educational Testing Service.

Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), *Measurement and personality assessment* (pp. 101–112). Elsevier Science Ltd.

Wright, B. D. (2001). Separation, reliability and skewed distributions: statistically different sample-independent levels of performance. *Rasch Measurement Transactions, 14*(4), 786. https://www.rasch.org/rmt/rmt144k.htm

Wright, B. D., & Douglas, G. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement,* 37(3), 47-60. https://doi.org/10.1177/001316447703700301

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* MESA Press.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of rasch measurement models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to rasch measurement: Theory, models and applications* (pp. 1-24). JAM Press.

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 29(1), 23-48. https://doi.org/10.1177/001316446902900102

Wright, B. D., & Stone, M. H. (1979). *Best test design.* MESA Press.

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wide Range, Inc.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation,* 31(2-3), 114-128. https://doi.org/10.1016/j.stueduc.2005.05.005

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: theory into practice*. Springer Nature Singapore Pte Ltd.

Wu, X., Wu, R., Chang, H.-H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology, 11*: 2230. https://doi.org/10.3389/fpsyg.2020.02230

Wyse, A. E. (2011). The similarity of bookmark cut scores with different response probability values. *Educational and Psychological Measurement, 71*(6), 963-985. https://doi.org/10.1177/0013164410395577

Wyse, A. E. (2013). Construct maps as a foundation for standard setting. *Measurement, 11*(4), 139-170. https://doi.org/10.1080/15366367.2013.850287

Wyse, A. E., & Babcock, B. (2020). It's not just Angoff: Misperceptions of hard and easy items in Bookmark-type ratings. *Educational Measurement: Issues and Practice, 39*(1), 22-29. https://doi.org/10.1111/emip.12315

UNIVERSITAS GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement,* 8(2), 125-145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada. https://www.ets.org/Media/Research/pdf/Yen_OPI_1987.pdf

Zapata-Rivera, D. (Ed.). (2018). *Score reporting research and applications*. Routledge.

Zeraati, M., & Alavi, N. M. (2014). Designing and validity evaluation of quality of nursing care scale in intensive care units. *Journal of Nursing Measurement, 22*(3), 461-471. http://dx.doi.org/10.1891/1061-3749.22.3.461

Zhu, Z. (2023). International comparative study of learning trajectories based on TIMSS 2019 G4 data on cognitive diagnostic models. *Frontiers in Psychology, 14*: 1241656. https://doi.org/10.3389/fpsyg.2023.1241656

Ziegler, L. A. (2014). *Reconceptualizing statistical literacy: developing an assessment for the modern introductory statistics course.* Doctoral dissertation, University of Minnesota

Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal, 17*(2), 161–178. https://doi.org/10.52041/serj.v17i2.164

UNIVERSITAS
GADJAH MADA
Pengembangan Instrumen dan Standard-Setting Literasi Statistika Untuk Mahasiswa Psikologi Jenjang Sarjana: Studi Berbasis Rasch Model
Muhammad Dwi Rifqi Kharisma Putra, Prof. Dr. Faturochman, M.A., Bahrul Hayat, M.A., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Zieky, M. J. (2012). So much has changed: an historical overview of setting cut scores. In G. J. Cizek (Ed.), *Setting performance standards: foundations, methods, and innovations* (2nd ed., pp. 15-32). Routledge.

Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive labs*. Pearson Education, Inc.

Zumbo, B. D. (2016). Standard-setting methodology: Establishing performance standards and setting cut-scores to assist score interpretation. *Applied Physiology, Nutrition, and Metabolism*, *41*, S74-S82. https://doi.org/10.1139/apnm-2015-0522

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment, 19*(2), 116–138. https://doi.org/10.1080/10627197.2014.903653

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15-25. https://doi.org/10.1111/j.1745-3992.2001.tb00059.x